

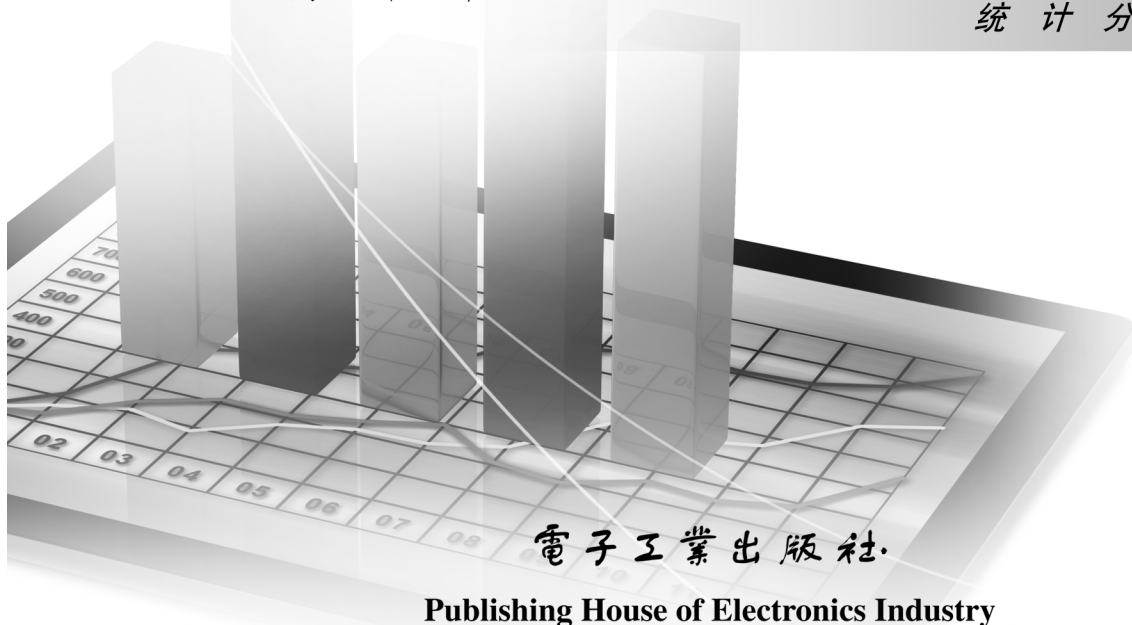
SAS

常用统计分析教程 (第2版)

◎ 胡良平 主编
◎ 高 辉 审校



统 计 分 析 系 列



電子工業出版社

Publishing House of Electronics Industry

北京 · BEIJING

内 容 简 介

本书内容丰富且新颖,适用面宽且可操作性强。涉及 SAS 软件基础和五种高级编程技术、统计设计中关键技术的 SAS 实现、定量与定性资料差异性和预测性分析。这些内容高质量、高效率地解决了实验设计、统计表达与描述、各种常用统计分析、现代回归分析、SAS 高级编程技术和 SAS 实现及结果解释等人们迫切需要解决却又十分棘手的问题。

本书第 1、2 篇共 7 章,介绍了 SAS 软件应用入门、SAS 语言基础、五种 SAS 高级编程技术,介绍了用 SAS 实现实验设计的关键技术(包括样本含量与检验效能估计、随机化和直接生成设计类型);第 3、4 篇共 8 章,对各种单因素和多因素设计下定量与定性结果进行差异性分析;第 5、6 篇共 16 章,对定量与定性结果提供了数十种预测性分析方法,包括定量和定性原因变量的判别分析。还有一些配套的辅助资料(实例数据等),放在华信教育资源网 www.hxedu.com.cn 上,便于读者查询。

本书适合需要学习和运用实验设计、统计表达与描述和统计分析及 SAS 实现的本科生、研究生、博士生、科研和管理工作者、临床医生和杂志编辑学习和使用。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究。

图书在版编目(CIP)数据

SAS 常用统计分析教程/胡良平主编. —2 版. —北京:电子工业出版社, 2015. 9

(统计分析系列)

ISBN 978-7-121-26831-1

I. ①S… II. ①胡… III. ①统计分析-应用软件-高等学校-教材 IV. ①C819

中国版本图书馆 CIP 数据核字(2015)第 176468 号

策划编辑:秦淑灵

责任编辑:郝黎明 特约编辑:张燕虹

印 刷:

装 订:

出版发行:电子工业出版社

北京市海淀区万寿路 173 信箱 邮编 100036

开 本:787×1092 1/16 印张:38.75 字数:992 千字

版 次:2010 年 6 月第 1 版

2015 年 9 月第 2 版

印 次:2015 年 9 月第 1 次印刷

印 数:3000 册 定价:75.00 元

凡所购买电子工业出版社图书有缺损问题,请向购买书店调换。若书店售缺,请与本社发行部联系,联系及邮购电话:(010)88254888。

质量投诉请发邮件至 zlt@phei.com.cn,盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线:(010)88258888。

编 委 会

主 编 胡良平

审 校 高 辉

副主编 柳伟伟 周诗国 胡纯严 郭 晋 李长平 葛 毅

编 委(以单位及姓氏笔画为序)

山西医科大学 余红梅 张岩波 郭东星

中山医科大学 王 霄 刘明华 张晋昕 张 熙 杨业春 薛允莲

中日友好医院 周诗国

天津医科大学 李长平

北京大学医学部 曹 波

军事医学科学院 王 琪 刘一松 孙日扬 吕辰龙 柳伟伟

胡 完 胡良平 胡纯严 郭辰仪 郭春雪

解放军卫生信息中心 葛 毅

解放军 95969 部队 高 辉

解放军第 273 医院 程瑞专

武汉大学 毛宗福 崔 丹

武警后勤学院 张 泽 武 佳

济南军区疾控中心 李子建

首都医科大学 刘惠刚 郭 晋 陶丽新

PPD 医药公司 毛 玮

第2版前言

在第1版于2010年6月由电子工业出版社正式出版后的2~3年间,此书经受住了市场的考验,得到了广大读者的褒奖、支持和鼓励;大约在2013年下半年,因市场上脱销了此书,故出版社征求笔者意见后,重印了数千册。最近,笔者在主讲全国广义综合评价统计学培训班期间,听见部分学员说此书在全国各大网店上缺货了。然而,早在2014年年底,负责此书出版的编辑就预见到了可能出现脱销现象,并敦促笔者抓紧时间修订。因为SAS软件内容更新很快,第1版主要基于SAS 6.1.3版本,而现在,SAS 9.2和SAS 9.3版本较之前增添了许多新功能。另外,第1版出版后,在使用过程中也发现了个别不妥和疏漏之处,有必要对第1版中的差错之处进行纠正;同时,有必要结合SAS软件的特点和实际使用统计学的需求,对全书内容和布局进行必要的修改、补充和调整。

在第2版中,增加了两篇内容,即第1篇“SAS软件及相关知识介绍”(第1章保留第1版的内容,第2章“SAS语言基础介绍”和第3章“SAS高级编程技术介绍”是新增的)和第2篇“统计设计中关键技术的SAS实现”(第4章“统计设计核心内容介绍”是必须交代的概念,第5~7章是实验设计中可用SAS实现的三个关键技术,即样本含量和检验效能估计、随机化和生成实验设计类型)。为了避免内容上的重复,删除了第1版中的第46章和第47章。值得一提的是,在第2版的第3章中,介绍了5种SAS高级编程技术:SAS ODS、SAS宏、SAS/SQL、SAS数组和SAS/IML。这些内容极其有用,掌握它们,标志着SAS用户SAS软件水平比较高,使用SAS的技能比较强。

除了上述有大幅度调整的内容之外,在第2版中,还对第1版中其他章节做了一些必要的修改和增减。例如,第2版的第8章,在第1版的第2章基础上增加了三节(特别是第8.5节,新增了实用价值很高的内容,即“成组设计一元定量资料三种特殊的比较——优效性、非劣效性和等效性t检验”)、修改了一节;第2版的第10章第10.11节,在第1版的第4章第4.11节的基础上补充了“具有一个重复测量因素的两因素设计的实例及SAS实现”,因为这是实际工作者最常使用的一种重复测量的设计类型;第2版的第13章,在第1版的第7章基础上增加了第15节,即“成组设计一元定性资料三种特殊的比较——优效性、非劣效性和等效性t检验”,这又是一个实用价值很高的内容;第2版的第16章和第19~22章,是在第1版的第10章和第13~16章基础上做了较大篇幅的补充,使之进一步完善和实用;第2版的第23章和第24章是由第1版的第17章扩展而成的,使条理更加清楚,便于读者学习和使用。另有些内容放于后续出版的《SAS高级统计分析教程(第2版)》中。

为方便读者查阅,本书还提供配套的辅助资料,包括实例数据等,可登录华信教育资源网 www.hxedu.com.cn 免费注册下载。

本书中有大量的程序语句、输出结果和计算公式,对于其中的变量,为了方便读者阅读,避免歧义,不再区分正、斜体,而是统一采用正体,特此说明。

在第2版即将出版之际，笔者情不自禁地要提及：本室2013级和2014级硕士研究生刘一松、胡完和郭春雪，他们不仅帮助完成了新增的第2~7章内容，还对全书做了认真的校对；最后，请允许笔者感谢直接和间接为此书第1版和第2版付出过辛勤劳动的所有同志和朋友！

由于笔者水平有限，书中难免会出现这样或那样的不妥，甚至错误之处，恳请广大读者不吝赐教，以便再版时修正。为便于与读者沟通和交流，给出笔者邮箱地址和有关网址：
lphu812@sina.com；www.statwd.com；www.huasitai.com。

主编 胡良平
于北京军事医学科学院
生物医学统计学咨询中心
2015年5月29日

第1版前言

众所周知,统计学内容非常丰富,学习和正确使用它的难度很大;SAS 软件功能非常强大,适用面极宽,SAS 语言又十分繁杂,学习和全面掌握其用法并非易事;显然,实际工作者要想在较短的时间内,学会用 SAS 软件方便快捷且正确地解决各种实验设计、统计表达与描述、常见和多元统计分析、现代回归分析、数据挖掘和基因表达谱分析等问题,几乎是天方夜谭,然而,本书却使其成为现实!

笔者有何灵丹妙药呢?“面向问题”的思维模式和写作手法是解决复杂问题并使其化繁为简、实用方便的“锦囊妙计”。本书各章针对拟解决的每个具体问题,首先给出“问题、数据和统计分析方法的选择”,接着,用编制好的 SAS 程序分析给定的资料,并给出程序修改指导、主要输出结果及其解释。本书正文内容共 8 篇 47 章。第 1 篇对定量结果进行差异性分析;第 2 篇对定性结果进行差异性分析;第 3 篇对定量结果进行预测性分析;第 4 篇对定性结果进行预测性分析;第 5 篇是多变量间相互与依赖关系分析;第 6 篇是变量或样品间亲疏关系或近似程度分析;第 7 篇是数据挖掘技术与基因表达谱分析简介;第 8 篇用编程法绘制统计图与实现实验设计。各章末尾均注明参编者详细名单。值得一提的是,本书中的所有计算基于 SAS 9.2 版本,少量 SAS 过程(特别是涉及 SAS/Genetics 模块)在 SAS 9.2 以前的环境下不能正常运行。

与本书配套光盘上的内容有:附录 1 是与 SAS 语言有关的内容简介,包括第 48 章 SAS 语句简介、第 49 章 SAS 过程简介、第 50 章 SAS 函数简介、第 51 章 SAS 宏简介、第 52 章 SAS ODS 简介、第 53 章 SAS SQL 简介、第 54 章 SAS 数组简介和第 55 章 SAS/IML 简介;附录 2 是四个非编程模块简介,包括第 56 章 SAS/ASSIST 模块用法简介、第 57 章 SAS/ANALYST 模块用法简介、第 58 章 SAS/INSIGHT 模块用法简介和第 59 章 SAS/ADX 模块用法简介;附录 3 是数据挖掘技术与基因表达谱分析,包括第 60 章数据挖掘的概念及常用统计分析技术、第 61 章基因表达谱的概念及数据分析技术和第 62 章生物信息学;附录 4 是各章实例和数据;附录 5 是各章 SAS 程序;附录 6 是各章 SAS 输出结果;附录 7 是各章计算原理和计算公式;附录 8 是各章参考文献;附录 9 是胡良平统计学专著与配套软件简介。

在本书即将出版之际,笔者真挚地感谢为本书做出过突出贡献的来自全国 10 多所高等院校的教授、副教授和青年学者,如山西医科大学张岩波、余红梅和郭东星教授,中山医科大学张晋昕教授,武汉大学毛宗福教授,哈尔滨医科大学李霞和张瑞杰教授,第三军医大学易东教授,首都医科大学郭秀花教授,北京大学医学部曹波副教授,北京邮电大学张晓航副教授等;还要感谢所有为本书付出过辛勤劳动的人们,他们的名字被写在本书的编委会名单中。特别是高辉、柳伟伟、周诗国、郭晋为本书的审阅工作付出了大量细致而又卓有成效的劳动,而本室 2009 级四名硕士研究生(鲍晓蕾、贾元杰、关雪、梦冰)也为本书的校对做

出了很多贡献。正是由于他们的积极参与、不懈努力和真心奉献，才使这部历时4年的专著能够问世！

由于笔者水平有限，书中难免会出现这样或那样的不妥，甚至错误之处，恳请广大读者不吝赐教，以便再版时修正。

主编 胡良平
于北京军事医学科学院
生物医学统计学咨询中心
2010年1月14日

目 录

第 1 篇 SAS 软件及相关知识介绍

| | | |
|-------|--------------------------------------|------|
| 第 1 章 | SAS 软件与 SAS 用法简介 | (1) |
| 1.1 | SAS 软件简介 | (1) |
| 1.1.1 | SAS 软件结构 | (1) |
| 1.1.2 | SAS 界面简介 | (1) |
| 1.1.3 | SAS 过程与 SAS 程序 | (2) |
| 1.1.4 | 运行 SAS 软件的两种常用方式 | (2) |
| 1.1.5 | SAS 程序结构 | (2) |
| 1.1.6 | 简单 SAS 程序中的 SAS 语句简介 | (3) |
| 1.1.7 | SAS 语言简介 | (5) |
| 1.1.8 | SAS 数据集简介 | (5) |
| 1.1.9 | 如何利用 SAS 帮助窗口 | (6) |
| 1.2 | SAS 用法简介 | (6) |
| 1.2.1 | 初学者学习 SAS 的快捷方式 | (6) |
| 1.2.2 | 实际运行 SAS | (7) |
| 1.2.3 | 从实验设计角度谈 SAS 用法 | (8) |
| 1.2.4 | 从资料录入角度谈 SAS 用法 | (8) |
| 1.2.5 | 从不同格式数据转换角度谈 SAS 用法 | (12) |
| 1.2.6 | 从资料表达角度谈 SAS 用法 | (13) |
| 1.2.7 | 从统计分析角度谈 SAS 用法 | (13) |
| 1.3 | 本章小结 | (14) |
| 第 2 章 | SAS 语言基础介绍 | (15) |
| 2.1 | SAS 数据步中常用 SAS 语句 | (15) |
| 2.1.1 | 数据获取语句 | (16) |
| 2.1.2 | 数据步文件管理语句 | (18) |
| 2.1.3 | SAS 变量操作语句 | (27) |
| 2.1.4 | SAS 观测值操作语句 | (32) |
| 2.1.5 | 数据步循环与控制语句 | (41) |
| 2.2 | SAS 过程步中常用 SAS 语句 | (46) |
| 2.3 | 可在 SAS 程序中任何地方出现的 SAS 语句——全程语句 | (48) |
| 2.3.1 | 全程数据存取语句 | (48) |
| 2.3.2 | 全程日志控制语句 | (49) |
| 2.3.3 | 全程环境控制语句 | (49) |

| | | |
|-------|--|------|
| 2.3.4 | 全局输出控制语句 | (49) |
| 2.3.5 | 全程序控制语句 | (50) |
| 2.4 | SAS 函数中的基础知识 | (51) |
| 2.4.1 | SAS 函数 | (51) |
| 2.4.2 | SAS 参数 | (51) |
| 2.4.3 | 函数值 | (52) |
| 2.4.4 | SAS 函数分类 | (52) |
| 2.4.5 | SAS 函数在使用中的注意事项 | (52) |
| 2.5 | 日期时间函数 | (53) |
| 2.5.1 | 日期时间函数简介 | (53) |
| 2.5.2 | 用 DATDIF 函数计算两个日期之间的天数 | (54) |
| 2.5.3 | 用 YRDIF 函数计算两个日期之间的年数 | (54) |
| 2.5.4 | 用 HOUR 和 MINUTE 函数计算当前时间 | (55) |
| 2.5.5 | 用 YEAR、QTR、MONTH 和 DAY 函数计算当前所处的年、季度、月份和日期 | (55) |
| 2.5.6 | 用 HOLIDAY 函数计算指定年份指定节日的日期 | (55) |
| 2.6 | 截取函数 | (56) |
| 2.6.1 | 截取函数简介 | (56) |
| 2.6.2 | 用 CEIL 函数求最小整数 | (56) |
| 2.6.3 | 用 FLOOR 函数求最大整数 | (56) |
| 2.6.4 | 用 INT 函数取整数部分 | (57) |
| 2.6.5 | 用 ROUND 函数按指定的精度取舍入值 | (57) |
| 2.6.6 | 用 TRUNC 函数求截取数值 | (57) |
| 2.7 | 分位数函数 | (57) |
| 2.7.1 | 分位数函数简介 | (57) |
| 2.7.2 | 用 CINV 函数计算卡方分布曲线下的 p 分位数 | (58) |
| 2.7.3 | 用 FINV 函数计算 F 分布曲线下的 p 分位数 | (58) |
| 2.7.4 | 用 PROBIT 函数计算标准正态分布曲线下的 p 分位数 | (59) |
| 2.7.5 | 用 TINV 函数计算 t 分布曲线下的 p 分位数 | (59) |
| 2.8 | 数学函数 | (60) |
| 2.8.1 | 数学函数简介 | (60) |
| 2.8.2 | 用 ABS 函数求绝对值 | (60) |
| 2.8.3 | 用 EXP 函数计算 e 的 x 次幂 | (61) |
| 2.8.4 | 用 LOG 函数计算以 e 为底的真数 x 的自然对数值 | (61) |
| 2.8.5 | 用 LOG10 函数计算以 10 为底的真数 x 的对数值 | (61) |
| 2.8.6 | 用 MOD 函数计算余数值 | (61) |
| 2.8.7 | 用 SQRT 函数计算平方根 | (62) |
| 2.8.8 | 用 SQRT 函数、FNONCT 函数和 FINV 函数计算 Ψ 值 | (62) |
| 2.8.9 | 用 CNONCT 函数和 CINV 函数计算 λ 值 | (63) |
| 2.9 | 概率函数 | (63) |
| 2.9.1 | 概率函数简介 | (63) |
| 2.9.2 | 用 PROBCHI 函数计算服从卡方分布的随机变量小于 x 的概率 | (64) |

| | | |
|--------|---|-------|
| 2.9.3 | 用 PROBF 函数计算服从 F 分布的随机变量小于 x 的概率 | (64) |
| 2.9.4 | 用 PROBNORM 函数计算标准正态分布曲线下的面积 | (64) |
| 2.9.5 | 用 PROBT 函数计算服从 t 分布的随机变量小于 x 的概率 | (65) |
| 2.9.6 | 用 PROBMC 函数计算 q 临界值 | (65) |
| 2.10 | 样本统计函数 | (67) |
| 2.10.1 | 样本统计函数简介 | (67) |
| 2.10.2 | 用 MEAN、MAX 与 MIN 函数计算算术均值、最大值与最小值 | (67) |
| 2.10.3 | 用 SUM、USS 与 CSS 函数计算和、未校正平方和与校正平方和 | (68) |
| 2.10.4 | 用 VAR、STD、STDERR 和 CV 函数计算方差、标准差、标准误与变异系数 | (68) |
| 2.10.5 | 用 SKEWNESS 和 KURTOSIS 函数计算偏度系数与峰度系数 | (68) |
| 2.10.6 | 用 NMISS 函数计算缺失值的个数 | (69) |
| 2.11 | 随机数函数 | (69) |
| 2.11.1 | 随机数函数简介 | (69) |
| 2.11.2 | 用 NORMAL 函数或 RANNOR 函数产生正态分布的随机数 | (69) |
| 2.11.3 | 用 UNIFORM 或 RANUNI 函数产生均匀分布的随机数 | (70) |
| 2.11.4 | 用 RANEXP 函数产生指数分布的随机数 | (71) |
| 2.11.5 | 用 RANBIN 函数产生二项分布的随机数 | (71) |
| 2.11.6 | 用 RANPOI 函数产生泊松分布的随机数 | (71) |
| 2.12 | SAS call 子程序 | (72) |
| 2.12.1 | 随机数子程序 | (72) |
| 2.12.2 | 其他子程序 | (72) |
| 2.12.3 | 随机数子程序的运用 | (72) |
| 第 3 章 | SAS 高级编程技术介绍 | (75) |
| 3.1 | SAS ODS 介绍 | (75) |
| 3.1.1 | 概述 | (75) |
| 3.1.2 | ODS 特点和常用输出目标 | (75) |
| 3.1.3 | 常用 ODS 语句 | (77) |
| 3.1.4 | SAS ODS 的应用 | (90) |
| 3.2 | SAS 宏介绍 | (100) |
| 3.2.1 | 概述 | (100) |
| 3.2.2 | 宏变量 | (101) |
| 3.2.3 | 宏与宏参数 | (108) |
| 3.2.4 | 宏的引用 | (110) |
| 3.2.5 | 常用宏语句和系统宏函数 | (116) |
| 3.2.6 | 宏与其他模块接口 | (126) |
| 3.3 | SAS SQL 介绍 | (128) |
| 3.3.1 | SQL 简介 | (128) |
| 3.3.2 | SQL 过程的语句介绍 | (129) |
| 3.4 | SAS 数组介绍 | (157) |
| 3.4.1 | 概述 | (157) |
| 3.4.2 | Array 语法格式 | (157) |

| | | |
|-------|--------------------|-------|
| 3.4.3 | 数组 Array 定义 | (158) |
| 3.4.4 | 数组 Array 初始化 | (160) |
| 3.4.5 | 数组引用 | (161) |
| 3.4.6 | 有关数组的 SAS 函数 | (163) |
| 3.5 | SAS/IML 介绍 | (164) |
| 3.5.1 | 概述 | (164) |
| 3.5.2 | 由矩阵标识创建矩阵 | (165) |
| 3.5.3 | 矩阵操作 | (167) |
| 3.5.4 | SAS/IML 编程语句 | (177) |
| 3.5.5 | IML 中常用函数 | (184) |
| 3.5.6 | IML 中数据集的操作 | (192) |

第 2 篇 统计设计中关键技术的 SAS 实现

| | | |
|-------|---------------------------|-------|
| 第 4 章 | 统计设计核心内容介绍 | (201) |
| 4.1 | 统计设计概述 | (201) |
| 4.1.1 | 统计设计类型 | (201) |
| 4.1.2 | 三类统计设计的共性 | (201) |
| 4.1.3 | 三类统计设计的个性 | (201) |
| 4.1.4 | 试验设计要点 | (202) |
| 4.1.5 | 临床试验设计要点 | (204) |
| 4.1.6 | 调查设计要点 | (206) |
| 4.2 | 设计类型概述 | (207) |
| 4.2.1 | 单因素设计 | (207) |
| 4.2.2 | 多因素设计 | (208) |
| 4.2.3 | 重复测量设计 | (212) |
| 4.3 | 比较类型概述 | (214) |
| 4.3.1 | 四种比较类型的概念 | (214) |
| 4.3.2 | 四种比较类型下检验假设及结论的正确陈述 | (215) |
| 4.3.3 | 合理选择临床试验的比较类型 | (216) |
| 4.4 | 样本含量与检验效能估计概述 | (217) |
| 4.4.1 | 样本含量估计的概念、意义与作用 | (217) |
| 4.4.2 | 检验效能估计的概念、意义与作用 | (218) |
| 4.5 | 随机化方法概述 | (218) |
| 4.5.1 | 随机化的概念 | (218) |
| 4.5.2 | 随机化的意义与作用 | (218) |
| 4.5.3 | 随机抽样方法 | (218) |
| 4.5.4 | 随机分组方法 | (219) |
| 4.6 | 本章小节 | (219) |
| 第 5 章 | 构建设计类型的 SAS 实现 | (220) |
| 5.1 | 常用标准多因素设计类型的列表格式 | (220) |

| | | |
|--------|------------------------------|-------|
| 5.1.1 | 随机区组设计 | (220) |
| 5.1.2 | 含一个协变量的随机区组设计 | (220) |
| 5.1.3 | 平衡不完全随机区组设计 | (221) |
| 5.1.4 | 拉丁方设计 | (221) |
| 5.1.5 | 交叉设计 | (221) |
| 5.1.6 | 无重复实验的双因素设计 | (222) |
| 5.1.7 | 嵌套设计 | (222) |
| 5.1.8 | 裂区设计 | (222) |
| 5.1.9 | 析因设计 | (223) |
| 5.1.10 | 含区组因素的析因设计 | (223) |
| 5.1.11 | 正交设计 | (224) |
| 5.1.12 | 均匀设计 | (225) |
| 5.1.13 | 重复测量设计 | (225) |
| 5.2 | 常用标准多因素设计类型的 SAS 输出格式 | (226) |
| 5.2.1 | 如何用 SAS 实现随机区组设计 | (226) |
| 5.2.2 | 如何用 SAS 实现平衡不完全区组设计 | (226) |
| 5.2.3 | 如何用 SAS 实现拉丁方设计 | (228) |
| 5.2.4 | 如何用 SAS 实现 2×2 交叉设计 | (228) |
| 5.2.5 | 如何用 SAS 实现 3×3 交叉设计 | (229) |
| 5.2.6 | 如何用 SAS 实现裂区设计 | (231) |
| 5.2.7 | 如何用 SAS 实现析因设计 | (232) |
| 5.2.8 | 如何用 SAS 实现含区组因素的析因设计 | (233) |
| 5.3 | 本章小结 | (234) |
| 第 6 章 | 样本含量与检验效能估计的 SAS 实现 | (235) |
| 6.1 | 估计样本含量与检验效能的前提条件 | (235) |
| 6.2 | 抽样调查中样本含量估计 | (236) |
| 6.2.1 | 估计总体均值时如何估计样本含量 | (236) |
| 6.2.2 | 估计总体率时如何估计样本含量 | (236) |
| 6.3 | 定量资料假设检验中样本含量与检验效能估计 | (237) |
| 6.3.1 | 单组、配对或交叉设计定量资料统计分析时样本含量估计 | (237) |
| 6.3.2 | 成组设计统计分析时样本含量估计 | (238) |
| 6.3.3 | 成组设计等效性检验时样本含量估计 | (238) |
| 6.3.4 | 成组设计非劣效或优效性检验时样本含量估计 | (239) |
| 6.3.5 | 单因素多水平设计定量资料方差分析时样本含量的估计 | (239) |
| 6.3.6 | 两因素析因设计定量资料方差分析时样本含量估计 | (240) |
| 6.3.7 | 简单直线相关或回归分析时样本含量的估计 | (242) |
| 6.3.8 | 单组、配对或交叉设计定量资料假设检验时检验效能的计算 | (243) |
| 6.3.9 | 成组设计均值差异性检验时检验效能的计算 | (243) |
| 6.3.10 | 成组设计均值等效性检验时检验效能的计算 | (243) |
| 6.3.11 | 成组设计均值非劣效或优效性检验时检验效能的计算 | (244) |
| 6.3.12 | 单因素多水平设计定量资料的方差分析时检验效能的计算 | (244) |

| | | |
|--------|---------------------------------|-------|
| 6.3.13 | 两因素析因设计定量资料方差分析时检验效能的计算 | (246) |
| 6.4 | 定性资料假设检验中样本含量与检验效能估计 | (248) |
| 6.4.1 | 单组设计率的检验时样本含量的估计 | (248) |
| 6.4.2 | 两样本频率比较时样本含量的估计 | (248) |
| 6.4.3 | 多个样本频率比较时样本含量的估计 | (249) |
| 6.4.4 | 单因素 2 水平设计定性资料等效性检验时检验效能的估计 | (249) |
| 6.4.5 | 单因素 2 水平设计定性资料非劣效或优效性检验时检验效能的估计 | (250) |
| 6.4.6 | 例数相等的两组样本频率比较时检验效能的计算 | (250) |
| 6.4.7 | 例数不相等的两组样本频率比较时检验效能的计算 | (251) |
| 6.4.8 | 单因素 2 水平设计定性资料等效性检验时检验效能的计算 | (251) |
| 6.4.9 | 单因素 2 水平设计定性资料非劣效或优效性检验时检验效能的计算 | (252) |
| 6.5 | 本章小结 | (252) |
| 第 7 章 | 随机化的 SAS 实现 | (253) |
| 7.1 | 常见随机抽样和随机分组的种类 | (253) |
| 7.2 | 调查研究中随机抽样的 SAS 实现 | (253) |
| 7.3 | 试验研究中随机分组的 SAS 实现 | (256) |
| 7.4 | 本章小结 | (263) |

第 3 篇 对定量结果进行差异性分析

| | | |
|-------|------------------------------------|-------|
| 第 8 章 | 单因素设计一元定量资料差异性分析 | (264) |
| 8.1 | 单组设计一元定量资料 t 检验与符号秩和检验 | (264) |
| 8.1.1 | 问题与数据 | (264) |
| 8.1.2 | 对数据结构的分析 | (264) |
| 8.1.3 | 分析目的与统计分析方法的选择 | (264) |
| 8.1.4 | SAS 程序中重要内容的说明 | (265) |
| 8.1.5 | 主要分析结果及解释 | (265) |
| 8.2 | 配对设计一元定量资料 t 检验与符号秩和检验 | (266) |
| 8.2.1 | 问题与数据 | (266) |
| 8.2.2 | 对数据结构的分析 | (267) |
| 8.2.3 | 分析目的与方法选择 | (267) |
| 8.2.4 | SAS 程序中重要内容的说明 | (267) |
| 8.2.5 | 主要分析结果及解释 | (267) |
| 8.3 | 成组设计一元定量资料 t 检验 | (268) |
| 8.3.1 | 问题与数据 | (268) |
| 8.3.2 | 对数据结构的分析 | (268) |
| 8.3.3 | 分析目的与方法选择 | (268) |
| 8.3.4 | SAS 程序中重要内容的说明 | (268) |
| 8.3.5 | 主要分析结果及解释 | (269) |
| 8.4 | 成组设计一元定量资料两种近似 t 检验和 Wilcoxon 秩和检验 | (270) |
| 8.4.1 | 问题与数据 | (270) |

| | | |
|-------|---|-------|
| 8.4.2 | 对数据结构的分析 | (270) |
| 8.4.3 | 分析目的与统计分析方法的选择 | (270) |
| 8.4.4 | SAS 程序中重要内容的说明 | (271) |
| 8.4.5 | 主要分析结果及解释 | (271) |
| 8.5 | 成组设计一元定量资料三种特殊的比较——优效性、非劣效性和等效性 t 检验 | (273) |
| 8.5.1 | 何为三种特殊的假设检验 | (273) |
| 8.5.2 | 成组设计一元定量资料优效性检验 | (273) |
| 8.5.3 | 成组设计一元定量资料非劣效性检验 | (274) |
| 8.5.4 | 成组设计一元定量资料等效性检验 | (275) |
| 8.6 | 单因素 $k(k \geq 3)$ 水平设计一元定量资料方差分析和两两比较 | (276) |
| 8.6.1 | 问题与数据 | (276) |
| 8.6.2 | 对数据结构的分析 | (276) |
| 8.6.3 | 分析目的与统计分析方法的选择 | (276) |
| 8.6.4 | SAS 程序中重要内容的说明 | (276) |
| 8.6.5 | 主要分析结果及解释 | (277) |
| 8.7 | 单因素 $k(k \geq 3)$ 水平设计定量资料一元协方差分析 | (278) |
| 8.7.1 | 问题与数据 | (278) |
| 8.7.2 | 对数据结构的分析 | (278) |
| 8.7.3 | 分析目的与统计分析方法的选择 | (279) |
| 8.7.4 | SAS 程序中重要内容的说明 | (279) |
| 8.7.5 | 主要分析结果及解释 | (279) |
| 8.8 | 单因素 $k(k \geq 3)$ 水平设计一元定量资料 Welch 近似方差分析和 Kruskal-Wallis 秩和检验及两两比较 | (280) |
| 8.8.1 | 问题与数据 | (280) |
| 8.8.2 | 对数据结构的分析 | (280) |
| 8.8.3 | 分析目的与统计分析方法的选择 | (280) |
| 8.8.4 | SAS 程序中重要内容的说明 | (280) |
| 8.8.5 | 主要分析结果及解释 | (281) |
| 8.9 | 本章小结 | (283) |
| 第 9 章 | 单因素设计一元生存资料差异性分析 | (284) |
| 9.1 | 单因素设计一元生存资料分析简介 | (284) |
| 9.2 | 生存资料统计描述 | (284) |
| 9.2.1 | 问题与数据 | (284) |
| 9.2.2 | 对数据结构的分析 | (285) |
| 9.2.3 | 分析目的与统计分析方法的选择 | (285) |
| 9.2.4 | SAS 程序 | (285) |
| 9.2.5 | 主要分析结果及解释 | (286) |
| 9.3 | 生存曲线比较 | (289) |
| 9.3.1 | 问题与数据 | (289) |
| 9.3.2 | 对数据结构的分析 | (289) |
| 9.3.3 | 分析目的与统计分析方法的选择 | (289) |

| | | |
|--------|---------------------------------|-------|
| 9.3.4 | SAS 程序 | (290) |
| 9.3.5 | 主要分析结果及解释 | (290) |
| 9.4 | 本章小结 | (291) |
| 第 10 章 | 多因素设计一元定量资料差异性分析 | (292) |
| 10.1 | 随机区组设计一元定量资料方差分析与 Friedman 秩和检验 | (292) |
| 10.1.1 | 问题与数据 | (292) |
| 10.1.2 | 对数据结构的分析 | (292) |
| 10.1.3 | 分析目的与统计分析方法的选择 | (292) |
| 10.1.4 | SAS 程序 | (292) |
| 10.1.5 | 主要分析结果及解释 | (294) |
| 10.2 | 双因素无重复实验设计一元定量资料方差分析 | (295) |
| 10.2.1 | 问题与数据 | (295) |
| 10.2.2 | 对数据结构的分析 | (296) |
| 10.2.3 | 分析目的与统计分析方法的选择 | (296) |
| 10.2.4 | SAS 程序 | (296) |
| 10.2.5 | 主要分析结果及解释 | (297) |
| 10.3 | 平衡不完全随机区组设计一元定量资料方差分析 | (298) |
| 10.3.1 | 问题与数据 | (298) |
| 10.3.2 | 对数据结构的分析 | (298) |
| 10.3.3 | 分析目的与统计分析方法的选择 | (298) |
| 10.3.4 | SAS 程序 | (298) |
| 10.3.5 | 主要分析结果及解释 | (299) |
| 10.4 | 拉丁方设计一元定量资料方差分析 | (300) |
| 10.4.1 | 问题与数据 | (300) |
| 10.4.2 | 对数据结构的分析 | (300) |
| 10.4.3 | 分析目的与统计分析方法的选择 | (301) |
| 10.4.4 | SAS 程序 | (301) |
| 10.4.5 | 主要分析结果及解释 | (301) |
| 10.5 | 二阶段交叉设计一元定量资料方差分析 | (302) |
| 10.5.1 | 问题与数据 | (302) |
| 10.5.2 | 对数据结构的分析 | (302) |
| 10.5.3 | 分析目的与统计分析方法的选择 | (302) |
| 10.5.4 | SAS 程序 | (303) |
| 10.5.5 | 主要分析结果及解释 | (303) |
| 10.6 | 析因设计一元定量资料方差分析 | (303) |
| 10.6.1 | 问题与数据 | (303) |
| 10.6.2 | 对数据结构的分析 | (304) |
| 10.6.3 | 分析目的与统计分析方法的选择 | (304) |
| 10.6.4 | SAS 程序 | (304) |
| 10.6.5 | 主要分析结果及解释 | (305) |
| 10.7 | 含区组因素的析因设计一元定量资料方差分析 | (306) |

| | | |
|---------|--------------------------------|-------|
| 10.7.1 | 问题与数据 | (306) |
| 10.7.2 | 对数据结构的分析 | (306) |
| 10.7.3 | 分析目的与统计分析方法的选择 | (306) |
| 10.7.4 | SAS 程序 | (307) |
| 10.7.5 | 主要分析结果及解释 | (307) |
| 10.8 | 嵌套设计一元定量资料方差分析 | (308) |
| 10.8.1 | 问题与数据 | (308) |
| 10.8.2 | 对数据结构的分析 | (308) |
| 10.8.3 | 分析目的与统计分析方法的选择 | (309) |
| 10.8.4 | SAS 程序 | (309) |
| 10.8.5 | 主要分析结果及解释 | (310) |
| 10.9 | 裂区设计一元定量资料方差分析 | (311) |
| 10.9.1 | 问题与数据 | (311) |
| 10.9.2 | 对数据结构的分析 | (312) |
| 10.9.3 | 分析目的与统计分析方法的选择 | (312) |
| 10.9.4 | SAS 程序 | (312) |
| 10.9.5 | 主要分析结果及解释 | (313) |
| 10.10 | 正交设计一元定量资料方差分析 | (314) |
| 10.10.1 | 问题与数据 | (314) |
| 10.10.2 | 对数据结构的分析 | (314) |
| 10.10.3 | 分析目的与统计分析方法的选择 | (315) |
| 10.10.4 | SAS 程序 | (315) |
| 10.10.5 | 主要分析结果及解释 | (316) |
| 10.11 | 重复测量设计一元定量资料方差分析 | (318) |
| 10.11.1 | 问题与数据 | (318) |
| 10.11.2 | 对数据结构的分析 | (319) |
| 10.11.3 | 分析目的与统计分析方法的选择 | (319) |
| 10.11.4 | SAS 程序 | (319) |
| 10.11.5 | 主要分析结果及解释 | (323) |
| 10.12 | 常见多因素实验设计一元定量资料协方差分析 | (330) |
| 10.12.1 | 问题与数据 | (330) |
| 10.12.2 | 对数据结构的分析 | (331) |
| 10.12.3 | 分析目的与统计分析方法的选择 | (332) |
| 10.12.4 | SAS 程序 | (332) |
| 10.12.5 | 主要分析结果及解释 | (334) |
| 10.13 | 多个单因素 2 水平设计定量资料 meta 分析 | (339) |
| 10.13.1 | 问题与数据 | (339) |
| 10.13.2 | 对数据结构的分析 | (339) |
| 10.13.3 | 分析目的与统计分析方法的选择 | (339) |
| 10.13.4 | SAS 程序 | (339) |
| 10.13.5 | 主要分析结果及解释 | (341) |
| 10.14 | 本章小结 | (341) |

| | | |
|--------|--|-------|
| 第 11 章 | 单因素设计多元定量资料差异性分析 | (342) |
| 11.1 | 问题、数据及统计分析方法的选择 | (342) |
| 11.1.1 | 问题与数据 | (342) |
| 11.1.2 | 对数据结构的分析 | (343) |
| 11.1.3 | 分析目的与统计分析方法选择 | (344) |
| 11.2 | 单因素设计定量资料多元方差和协方差分析 | (345) |
| 11.2.1 | 对例 11-1 资料进行单组设计定量资料二元方差分析 | (345) |
| 11.2.2 | 对例 11-2 资料进行配对设计定量资料二元方差分析 | (345) |
| 11.2.3 | 对例 11-3 资料进行单因素 2 水平设计定量资料三元方差分析 | (346) |
| 11.2.4 | 对例 11-4 资料进行单因素 3 水平设计定量资料二元方差分析 | (347) |
| 11.2.5 | 对例 11-5 资料进行单因素 2 水平设计二元定量资料的一元协方差分析 | (349) |
| 11.2.6 | 对例 11-6 资料进行单因素 2 水平设计二元定量资料的二元协方差分析 | (351) |
| 11.3 | 本章小结 | (354) |
| 第 12 章 | 多因素设计多元定量资料差异性分析 | (355) |
| 12.1 | 问题、数据及统计分析方法的选择 | (355) |
| 12.1.1 | 问题与数据 | (355) |
| 12.1.2 | 对数据结构的分析 | (358) |
| 12.1.3 | 分析目的与统计分析方法的选择 | (359) |
| 12.2 | 多因素设计定量资料多元方差和协方差分析 | (360) |
| 12.2.1 | 对例 12-1 资料进行随机区组设计定量资料三元方差分析 | (360) |
| 12.2.2 | 对例 12-2 资料进行两因素析因设计定量资料三元方差分析 | (361) |
| 12.2.3 | 对例 12-3 资料进行含区组因素析因设计定量资料四元方差分析 | (362) |
| 12.2.4 | 对例 12-4 资料进行正交设计定量资料三元方差分析 | (363) |
| 12.2.5 | 对例 12-5 资料进行具有一个重复测量的两因素设计定量资料二元方差分析 | (365) |
| 12.2.6 | 对例 12-6 资料进行两因素析因设计五元定量资料的二元协方差分析 | (366) |
| 12.3 | 本章小结 | (369) |

第 4 篇 对定性结果进行差异性分析

| | | |
|--------|------------------------|-------|
| 第 13 章 | 单因素设计一元定性资料差异性分析 | (370) |
| 13.1 | 单组设计一维表资料统计分析 | (370) |
| 13.1.1 | 问题与数据 | (370) |
| 13.1.2 | 对数据结构的分析 | (370) |
| 13.1.3 | 分析目的与统计分析方法的选择 | (370) |
| 13.1.4 | SAS 程序中重要内容的说明 | (370) |
| 13.1.5 | 主要分析结果及解释 | (371) |
| 13.2 | 配对设计四格表资料统计分析 | (371) |
| 13.2.1 | 问题与数据 | (371) |
| 13.2.2 | 对数据结构的分析 | (371) |
| 13.2.3 | 分析目的与统计分析方法的选择 | (371) |
| 13.2.4 | SAS 程序中重要内容的说明 | (372) |

| | | |
|--------|---------------------------------------|-------|
| 13.2.5 | 主要分析结果及解释 | (372) |
| 13.3 | 配对设计扩大形式的方表资料统计分析 | (373) |
| 13.3.1 | 问题与数据 | (373) |
| 13.3.2 | 对数据结构的分析 | (373) |
| 13.3.3 | 分析目的与统计分析方法的选择 | (373) |
| 13.3.4 | SAS 程序中重要内容的说明 | (373) |
| 13.3.5 | 主要分析结果及解释 | (374) |
| 13.4 | 成组设计横断面研究四格表资料统计分析 | (374) |
| 13.4.1 | 问题与数据 | (374) |
| 13.4.2 | 对数据结构的分析 | (374) |
| 13.4.3 | 分析目的与统计分析方法的选择 | (375) |
| 13.4.4 | SAS 程序中重要内容的说明 | (375) |
| 13.4.5 | 主要分析结果及解释 | (375) |
| 13.5 | 成组设计队列研究四格表资料统计分析 | (376) |
| 13.5.1 | 问题与数据 | (376) |
| 13.5.2 | 对数据结构的分析 | (376) |
| 13.5.3 | 分析目的与统计分析方法的选择 | (376) |
| 13.5.4 | SAS 程序中重要内容的说明 | (376) |
| 13.5.5 | 主要分析结果及解释 | (376) |
| 13.6 | 成组设计病例对照研究四格表资料统计分析 | (377) |
| 13.6.1 | 问题与数据 | (377) |
| 13.6.2 | 对数据结构的分析 | (378) |
| 13.6.3 | 分析目的与统计分析方法的选择 | (378) |
| 13.6.4 | SAS 程序中重要内容的说明 | (378) |
| 13.6.5 | 主要分析结果及解释 | (378) |
| 13.7 | 成组设计结果变量为多值有序变量的 $2 \times C$ 表资料统计分析 | (379) |
| 13.7.1 | 问题与数据 | (379) |
| 13.7.2 | 对数据结构的分析 | (379) |
| 13.7.3 | 分析目的与统计分析方法的选择 | (379) |
| 13.7.4 | SAS 程序中重要内容的说明 | (379) |
| 13.7.5 | 主要分析结果及解释 | (380) |
| 13.8 | 成组设计结果变量为多值名义变量的 $2 \times C$ 表资料统计分析 | (380) |
| 13.8.1 | 问题与数据 | (380) |
| 13.8.2 | 对数据结构的分析 | (380) |
| 13.8.3 | 分析目的与统计分析方法的选择 | (380) |
| 13.8.4 | SAS 程序中重要内容的说明 | (380) |
| 13.8.5 | 主要分析结果及解释 | (381) |
| 13.9 | 单因素多水平设计无序原因变量 $R \times 2$ 表资料统计分析 | (381) |
| 13.9.1 | 问题与数据 | (381) |
| 13.9.2 | 对数据结构的分析 | (381) |
| 13.9.3 | 分析目的与统计分析方法的选择 | (382) |
| 13.9.4 | SAS 程序中重要内容的说明 | (382) |

| | | |
|---------|--|-------|
| 13.9.5 | 主要分析结果及解释 | (382) |
| 13.10 | 单因素多水平设计有序原因变量 $R \times 2$ 表资料统计分析 | (382) |
| 13.10.1 | 问题与数据 | (382) |
| 13.10.2 | 对数据结构的分析 | (382) |
| 13.10.3 | 分析目的与统计分析方法的选择 | (383) |
| 13.10.4 | SAS 程序中重要内容的说明 | (383) |
| 13.10.5 | 主要分析结果及解释 | (383) |
| 13.11 | 单因素多水平设计双向无序 $R \times C$ 表资料统计分析 | (383) |
| 13.11.1 | 问题与数据 | (383) |
| 13.11.2 | 对数据结构的分析 | (384) |
| 13.11.3 | 分析目的与统计分析方法的选择 | (384) |
| 13.11.4 | SAS 程序中重要内容的说明 | (384) |
| 13.11.5 | 主要分析结果及解释 | (384) |
| 13.12 | 单因素多水平设计有序结果变量 $R \times C$ 表资料统计分析 | (385) |
| 13.12.1 | 问题与数据 | (385) |
| 13.12.2 | 对数据结构的分析 | (385) |
| 13.12.3 | 分析目的与统计分析方法的选择 | (385) |
| 13.12.4 | SAS 程序中重要内容的说明 | (385) |
| 13.12.5 | 主要分析结果及解释 | (385) |
| 13.13 | 单因素多水平设计双向有序 $R \times C$ 表资料统计分析 | (386) |
| 13.13.1 | 问题与数据 | (386) |
| 13.13.2 | 对数据结构的分析 | (386) |
| 13.13.3 | 分析目的与统计分析方法的选择 | (386) |
| 13.13.4 | SAS 程序中重要内容的说明 | (386) |
| 13.13.5 | 主要分析结果及解释 | (387) |
| 13.14 | 数据库形式表达资料的统计分析 | (387) |
| 13.15 | 成组设计一元定性资料三种特殊的比较——优效性、非劣效性和等效性 t 检验 | (388) |
| 13.15.1 | 成组设计一元定性资料三种特殊的比较概述 | (388) |
| 13.15.2 | 优效性检验 | (389) |
| 13.15.3 | 非劣效性检验 | (389) |
| 13.15.4 | 等效性检验 | (390) |
| 13.16 | 本章小结 | (391) |
| 第 14 章 | 多因素设计一元定性资料差异性分析 | (392) |
| 14.1 | 用加权 χ^2 检验处理结果变量为二值变量的高维列联表资料 | (392) |
| 14.1.1 | 问题与数据 | (392) |
| 14.1.2 | 对数据结构的分析 | (392) |
| 14.1.3 | 分析目的与统计分析方法的选择 | (392) |
| 14.1.4 | SAS 程序中重要内容的说明 | (393) |
| 14.1.5 | 主要分析结果及解释 | (393) |

| | | |
|--------|---------------------------------------|-------|
| 14.2 | 用 CMH χ^2 检验处理结果变量具有三种性质的高维列联表资料 | (393) |
| 14.2.1 | 问题与数据 | (393) |
| 14.2.2 | 对数据结构的分析 | (394) |
| 14.2.3 | 分析目的与统计分析方法的选择 | (394) |
| 14.2.4 | SAS 程序中重要内容的说明 | (395) |
| 14.2.5 | 主要分析结果及解释 | (396) |
| 14.3 | 用 meta 分析分别合并处理多个成组设计定性资料 | (397) |
| 14.3.1 | 问题与数据 | (397) |
| 14.3.2 | 对数据结构的分析 | (398) |
| 14.3.3 | 分析目的与统计分析方法的选择 | (399) |
| 14.3.4 | SAS 程序中重要内容的说明 | (399) |
| 14.3.5 | 主要分析结果及解释 | (403) |
| 14.4 | 用 ROC 方法分析诊断试验资料 | (404) |
| 14.4.1 | 问题与数据 | (404) |
| 14.4.2 | 对数据结构的分析 | (405) |
| 14.4.3 | 分析目的与统计分析方法的选择 | (405) |
| 14.4.4 | SAS 程序中重要内容的说明 | (405) |
| 14.4.5 | 主要分析结果及解释 | (408) |
| 14.5 | 本章小结 | (410) |
| 第 15 章 | 多因素设计一元定性资料的对数线性模型分析 | (411) |
| 15.1 | 问题、数据及统计分析方法的选择 | (411) |
| 15.1.1 | 问题与数据 | (411) |
| 15.1.2 | 对数据结构的分析 | (412) |
| 15.1.3 | 分析目的与统计分析方法的选择 | (412) |
| 15.2 | 用对数线性模型分析列联表资料 | (412) |
| 15.2.1 | 对数线性模型简介 | (412) |
| 15.2.2 | 用 SAS 分析例 15-1 资料 | (412) |
| 15.2.3 | 用 SAS 分析例 15-2 资料 | (417) |
| 15.3 | 本章小结 | (418) |

第 5 篇 对定量结果进行预测性分析

| | | |
|--------|-----------------|-------|
| 第 16 章 | 两变量简单线性相关与回归分析 | (419) |
| 16.1 | 问题、数据及统计分析方法的选择 | (419) |
| 16.1.1 | 问题与数据 | (419) |
| 16.1.2 | 对数据结构的分析 | (420) |
| 16.1.3 | 分析目的与统计分析方法的选择 | (421) |
| 16.1.4 | 统计分析方法简介 | (422) |
| 16.2 | Pearson 线性相关分析 | (423) |
| 16.2.1 | SAS 程序中重要内容的说明 | (423) |
| 16.2.2 | 主要分析结果及解释 | (423) |

| | | |
|--------|-------------------------------------|-------|
| 16.3 | Spearman 秩相关分析 | (424) |
| 16.3.1 | SAS 程序中重要内容的说明 | (424) |
| 16.3.2 | 主要分析结果及解释 | (424) |
| 16.4 | Kendall's Tau-b 相关分析 | (425) |
| 16.4.1 | SAS 程序中重要内容的说明 | (425) |
| 16.4.2 | 主要分析结果及解释 | (425) |
| 16.5 | 简单线性回归分析 | (426) |
| 16.5.1 | 对例 16-4 资料的分析 | (426) |
| 16.5.2 | 对例 16-5 资料的分析 | (427) |
| 16.6 | 常用于估计 LD50 的加权线性回归分析 | (430) |
| 16.6.1 | SAS 程序中重要内容的说明 | (430) |
| 16.6.2 | 主要分析结果及解释 | (430) |
| 16.6.3 | 用于比较 LD50 和斜率的 SAS 程序中重要内容的说明 | (432) |
| 16.6.4 | 两两比较的主要分析结果及解释 | (433) |
| 16.7 | 本章小结 | (434) |
| 第 17 章 | 两变量可直线化曲线回归分析 | (435) |
| 17.1 | 问题、数据及统计分析方法的选择 | (435) |
| 17.1.1 | 问题与数据 | (435) |
| 17.1.2 | 对数据结构的分析 | (435) |
| 17.1.3 | 分析目的与统计分析方法的选择 | (435) |
| 17.2 | 对数函数、幂函数和双曲函数曲线回归分析 | (436) |
| 17.2.1 | SAS 程序中重要内容的说明 | (436) |
| 17.2.2 | 主要分析结果及解释 | (437) |
| 17.3 | 指数函数曲线回归分析 | (440) |
| 17.3.1 | SAS 程序中重要内容的说明 | (440) |
| 17.3.2 | 主要分析结果及解释 | (441) |
| 17.4 | logistic 函数曲线回归分析 | (442) |
| 17.4.1 | SAS 程序中重要内容的说明 | (442) |
| 17.4.2 | 主要分析结果及解释 | (443) |
| 17.5 | 本章小结 | (444) |
| 第 18 章 | 各种复杂曲线回归分析 | (445) |
| 18.1 | 多项式曲线回归分析 | (445) |
| 18.1.1 | 问题与数据 | (445) |
| 18.1.2 | 分析目的与统计分析方法的选择 | (445) |
| 18.1.3 | SAS 程序 | (445) |
| 18.1.4 | 主要分析结果及解释 | (446) |
| 18.2 | logistic 曲线回归分析 | (447) |
| 18.2.1 | 问题与数据 | (447) |
| 18.2.2 | 分析目的与统计分析方法的选择 | (447) |
| 18.2.3 | SAS 程序 | (447) |

| | | |
|--------|---------------------|-------|
| 18.2.4 | 主要分析结果及解释 | (448) |
| 18.3 | Gompertz 曲线回归分析 | (449) |
| 18.3.1 | 问题与数据 | (449) |
| 18.3.2 | 分析目的与统计分析方法的选择 | (449) |
| 18.3.3 | SAS 程序 | (450) |
| 18.3.4 | 主要分析结果及解释 | (450) |
| 18.4 | 二项型指数曲线回归分析 | (452) |
| 18.4.1 | 问题与数据 | (452) |
| 18.4.2 | 分析目的与统计分析方法的选择 | (452) |
| 18.4.3 | SAS 程序 | (452) |
| 18.4.4 | 主要分析结果及解释 | (454) |
| 18.5 | 三项型指数曲线回归分析 | (456) |
| 18.5.1 | 问题与数据 | (456) |
| 18.5.2 | 分析目的与统计分析方法的选择 | (457) |
| 18.5.3 | SAS 程序 | (457) |
| 18.5.4 | 主要分析结果及解释 | (459) |
| 18.6 | 本章小结 | (462) |
| 第 19 章 | 多重线性回归分析 | (463) |
| 19.1 | 问题、数据及统计分析方法的选择 | (463) |
| 19.1.1 | 问题与数据 | (463) |
| 19.1.2 | 对数据结构的分析 | (463) |
| 19.1.3 | 分析目的与统计分析方法的选择 | (463) |
| 19.2 | 多重线性回归分析概述 | (464) |
| 19.3 | 产生哑变量和派生变量的方法 | (465) |
| 19.3.1 | 如何产生哑变量 | (465) |
| 19.3.2 | 如何产生派生变量 | (466) |
| 19.4 | 自变量各种筛选方法介绍 | (466) |
| 19.4.1 | 筛选自变量的必要性 | (466) |
| 19.4.2 | 筛选自变量的方法 | (466) |
| 19.4.3 | 如何在多个多重线性回归方程中选择最佳者 | (468) |
| 19.5 | 自变量各种共线性诊断方法介绍 | (468) |
| 19.5.1 | 共线性的概念 | (468) |
| 19.5.2 | 共线性的诊断 | (468) |
| 19.6 | 各种异常点诊断方法介绍 | (469) |
| 19.6.1 | 异常点的概念 | (469) |
| 19.6.2 | 异常点的诊断 | (469) |
| 19.7 | 自变量作用大小的评价 | (469) |
| 19.8 | 多个多重回归方程优劣的评价 | (470) |
| 19.9 | 多重线性回归分析的 SAS 实现 | (470) |
| 19.9.1 | SAS 程序及说明 | (470) |
| 19.9.2 | 主要分析结果及解释 | (471) |

| | | |
|---------------|-----------------------------|--------------|
| 19.10 | REG 过程语法简介 | (474) |
| 19.11 | 本章小结 | (476) |
| 第 20 章 | 主成分回归分析 | (477) |
| 20.1 | 问题、数据及统计分析方法的选择 | (477) |
| 20.1.1 | 问题与数据 | (477) |
| 20.1.2 | 对数据结构的分析 | (478) |
| 20.1.3 | 分析目的及统计分析方法的选择 | (478) |
| 20.2 | 主成分回归分析应用场合及关键技术 | (478) |
| 20.3 | 主成分回归分析 | (479) |
| 20.3.1 | 对例 20-1 资料进行主成分回归分析 | (479) |
| 20.3.2 | 对例 20-2 数据进行分析 | (483) |
| 20.4 | 本章小结 | (487) |
| 第 21 章 | 用 SAS 实现岭回归分析 | (488) |
| 21.1 | 问题、数据及统计分析方法的选择 | (488) |
| 21.1.1 | 问题与数据 | (488) |
| 21.1.2 | 对数据结构的分析 | (488) |
| 21.1.3 | 分析目的与统计分析方法的选择 | (489) |
| 21.2 | 岭回归分析应用场合及关键技术 | (489) |
| 21.3 | 岭回归分析 | (490) |
| 21.3.1 | 进行多重线性回归分析并进行共线性诊断 | (490) |
| 21.3.2 | 进行岭回归分析 | (491) |
| 21.4 | 与岭回归分析有关的 SAS 语句说明 | (494) |
| 21.5 | 本章小结 | (494) |
| 第 22 章 | Poisson 回归分析 | (495) |
| 22.1 | 问题、数据及统计分析方法的选择 | (495) |
| 22.1.1 | 问题与数据 | (495) |
| 22.1.2 | 对数据结构的分析 | (496) |
| 22.1.3 | 分析目的与统计分析方法的选择 | (496) |
| 22.2 | Poisson 回归分析应用场合及关键技术 | (497) |
| 22.2.1 | Poisson 回归模型简介 | (497) |
| 22.2.2 | Poisson 回归定义 | (497) |
| 22.2.3 | Possion 回归的参数估计 | (498) |
| 22.2.4 | Possion 回归估计系数的假设检验 | (498) |
| 22.2.5 | Possion 拟合优度检验 | (498) |
| 22.3 | Poisson 回归分析的 SAS 实现 | (499) |
| 22.3.1 | 对例 22-1 资料进行分析 | (499) |
| 22.3.2 | 对例 22-2 资料进行分析 | (502) |
| 22.4 | 本章小结 | (504) |
| 第 23 章 | 负二项回归分析 | (505) |
| 23.1 | 问题、数据及统计分析方法的选择 | (505) |

| | | |
|---------------|--------------------------|--------------|
| 23.1.1 | 问题与数据 | (505) |
| 23.1.2 | 对数据结构的分析 | (506) |
| 23.1.3 | 分析目的与统计分析方法的选择 | (506) |
| 23.2 | 负二项回归分析应用场合及关键技术 | (506) |
| 23.3 | 负二项回归分析的 SAS 实现 | (508) |
| 23.3.1 | SAS 程序及说明 | (508) |
| 23.3.2 | 主要分析结果及解释 | (508) |
| 23.4 | GENMOD 过程及 COUNTREG 过程简介 | (511) |
| 23.5 | 本章小结 | (513) |
| 第 24 章 | Probit 回归分析 | (514) |
| 24.1 | 问题、数据及统计分析方法的选择 | (514) |
| 24.1.1 | 问题与数据 | (514) |
| 24.1.2 | 对数据结构的分析 | (515) |
| 24.1.3 | 分析目的与统计分析方法的选择 | (515) |
| 24.2 | Probit 回归分析应用场合及关键技术 | (515) |
| 24.2.1 | 适于进行 Probit 回归分析的资料表达形式 | (515) |
| 24.2.2 | Probit 回归模型简介 | (516) |
| 24.3 | Probit 回归分析的 SAS 实现 | (516) |
| 24.3.1 | 对例 24-1 资料进行 Probit 回归分析 | (516) |
| 24.3.2 | 对例 24-2 资料进行 Probit 回归分析 | (518) |
| 24.4 | PROBIT 过程语法简介 | (523) |
| 24.5 | 本章小结 | (526) |
| 第 25 章 | 生存资料 Cox 模型回归分析 | (527) |
| 25.1 | 实例 | (527) |
| 25.1.1 | 问题与数据 | (527) |
| 25.1.2 | 对数据结构的分析 | (527) |
| 25.1.3 | 分析目的与统计分析方法的选择 | (527) |
| 25.2 | 生存资料 Cox 模型回归分析简介 | (528) |
| 25.3 | 生存资料 Cox 模型回归分析的 SAS 实现 | (528) |
| 25.3.1 | SAS 程序 | (528) |
| 25.3.2 | 主要分析结果及解释 | (530) |
| 25.4 | 本章小结 | (532) |
| 第 26 章 | 生存资料参数模型回归分析 | (533) |
| 26.1 | 实例 | (533) |
| 26.1.1 | 问题与数据 | (533) |
| 26.1.2 | 对数据结构的分析 | (533) |
| 26.1.3 | 分析目的与统计分析方法的选择 | (533) |
| 26.2 | 生存资料参数模型回归分析简介 | (534) |
| 26.3 | 生存资料参数模型回归分析的 SAS 实现 | (534) |
| 26.3.1 | SAS 程序 | (534) |

| | | |
|--------|--------------|-------|
| 26.3.2 | 主要分析结果及解释 | (535) |
| 26.4 | LIFEREG 过程简介 | (538) |
| 26.5 | 本章小结 | (539) |
| 第 27 章 | 时间序列分析 | (540) |
| 27.1 | 时间序列分析简介 | (540) |
| 27.2 | 指数平滑法 | (540) |
| 27.2.1 | 指数平滑法简介 | (540) |
| 27.2.2 | 应用实例 | (541) |
| 27.2.3 | SAS 程序 | (541) |
| 27.2.4 | 主要分析结果及解释 | (541) |
| 27.3 | ARIMA 模型 | (543) |
| 27.3.1 | ARIMA 模型简介 | (543) |
| 27.3.2 | 应用实例 | (543) |
| 27.3.3 | SAS 程序 | (543) |
| 27.3.4 | 主要分析结果及解释 | (544) |
| 27.4 | 谱分析 | (547) |
| 27.4.1 | 谱分析简介 | (547) |
| 27.4.2 | 应用实例 | (548) |
| 27.4.3 | SAS 程序 | (548) |
| 27.4.4 | 主要分析结果及解释 | (549) |
| 27.5 | X12 方法 | (551) |
| 27.5.1 | X12 过程简介 | (551) |
| 27.5.2 | 应用实例 | (551) |
| 27.5.3 | SAS 程序 | (551) |
| 27.5.4 | 摘录主要分析结果 | (552) |
| 27.6 | 本章小结 | (556) |

第 6 篇 对定性结果进行预测性分析

| | | |
|--------|---------------------------|-------|
| 第 28 章 | 非配对设计定性资料多重 logistic 回归分析 | (557) |
| 28.1 | 问题、数据及统计分析方法的选择 | (557) |
| 28.1.1 | 问题与数据 | (557) |
| 28.1.2 | 对数据结构的分析 | (558) |
| 28.1.3 | 分析目的与统计分析方法的选择 | (559) |
| 28.1.4 | logistic 回归分析简介 | (559) |
| 28.2 | 二值变量的多重 logistic 回归分析 | (559) |
| 28.2.1 | SAS 程序及说明 | (559) |
| 28.2.2 | 主要分析结果及结论 | (560) |
| 28.3 | 多值有序变量的多重 logistic 回归分析 | (562) |
| 28.3.1 | SAS 程序及说明 | (562) |
| 28.3.2 | 主要分析结果及结论 | (563) |

| | | |
|--------|-------------------------------|-------|
| 28.4 | 多值名义变量的多重 logistic 回归分析 | (565) |
| 28.4.1 | SAS 程序及说明 | (565) |
| 28.4.2 | 主要分析结果及结论 | (565) |
| 28.5 | 本章小结 | (567) |
| 第 29 章 | 配对设计定性资料多重 logistic 回归分析 | (568) |
| 29.1 | 问题、数据及统计分析方法的选择 | (568) |
| 29.1.1 | 问题与数据 | (568) |
| 29.1.2 | 对数据结构的分析 | (569) |
| 29.1.3 | 分析目的与统计分析方法的选择 | (569) |
| 29.1.4 | 条件 logistic 回归分析简介 | (569) |
| 29.2 | 1:1 配对设计定性资料的多重 logistic 回归分析 | (569) |
| 29.2.1 | SAS 程序及说明 | (569) |
| 29.2.2 | 主要分析结果及解释 | (570) |
| 29.3 | m:n 配对设计定性资料的多重 logistic 回归分析 | (572) |
| 29.3.1 | SAS 程序及说明 | (572) |
| 29.3.2 | 主要分析结果及解释 | (572) |
| 29.4 | 本章小结 | (574) |
| 第 30 章 | 原因变量为定量变量的判别分析 | (575) |
| 30.1 | 实例 | (575) |
| 30.1.1 | 问题与数据 | (575) |
| 30.1.2 | 对数据结构的分析 | (576) |
| 30.1.3 | 分析目的与统计分析方法的选择 | (576) |
| 30.2 | 原因变量为定量变量的判别分析简介 | (576) |
| 30.3 | 原因变量为定量变量的判别分析 | (577) |
| 30.3.1 | SAS 程序 | (577) |
| 30.3.2 | 主要分析结果及解释 | (579) |
| 30.4 | 本章小结 | (588) |
| 第 31 章 | 原因变量为定性变量的判别分析 | (589) |
| 31.1 | 实例 | (589) |
| 31.1.1 | 问题与数据 | (589) |
| 31.1.2 | 对数据结构的分析 | (590) |
| 31.1.3 | 分析目的与统计分析方法的选择 | (590) |
| 31.2 | 原因变量为定性变量的判别分析简介 | (590) |
| 31.3 | 原因变量为定性变量的判别分析 | (590) |
| 31.3.1 | SAS 程序 | (590) |
| 31.3.2 | 主要分析结果及解释 | (591) |
| 31.4 | 本章小结 | (592) |

第 1 篇 SAS 软件及相关知识介绍

第 1 章 SAS 软件与 SAS 用法简介

由于 SAS 软件产品很多,不同产品用法不尽相同。本章仅围绕与统计学内容密切相关的方面,扼要介绍 SAS 软件及与使用 SAS 有关的基本概念和用法。本书中绝大部分内容都属于详细介绍如何用 SAS 软件实现各种统计分析及其结果解释,本章仅是这些详细内容的一个缩影。确切地说,本章重点是介绍全书中其他章节不做详细介绍的与统计学应用密切相关的一些“基本常识”和“小知识点”,为读者顺利学习其他各章做一点铺垫。

1.1 SAS 软件简介

1.1.1 SAS 软件结构

SAS 软件采取模块式结构,每个模块可被称为一个 SAS 产品。有些 SAS 产品中仅有具体的 SAS 过程(即编译后的 SAS 程序),如 SAS/STAT;有些 SAS 产品中不仅有 SAS 过程,还有其他内容,如 SAS 语言、SAS 窗口、SAS 宏、SAS SQL 等,见 SAS/BASE;有些 SAS 产品本身就是一个功能比较齐全的软件,可以完成一系列相关的功能,如 SAS/ASSIST 模块、SAS/ANALYST 模块和 SAS/INSIGHT 模块等;还有些 SAS 产品是其他 SAS 产品的开发工具,如 SAS/AF 等。

1.1.2 SAS 界面简介

不同版本的 SAS 软件,其界面不尽相同。比较新的 SAS 软件(SAS 9.3)的主界面如图 1-1 所示。

由图 1-1 可知,在主界面上,顶部有两行,分别为“菜单栏”和“工具栏”,左边有一个“SAS 资源管理器”窗口,与它重叠的一个窗口叫“结果”窗口;右边有两个窗口,上方的窗口叫“日志”窗口,下方的窗口叫“程序编辑器”窗口,通过顶部菜单条中的“窗口”选项,可以切换到其他窗口,如“输出”窗口。

可以说,基本的 SAS 窗口有“SAS 资源管理器”、“结果”、“程序编辑”、“日志”和“输出”窗口,但另外还有 30 多个窗口可供用户处理打印和微调 SAS 会话之类的操作。这些窗口的名称和窗口命令的详细列表从略。若用户想获得此列表,可通过下面的方法实现:先选中主界面第 2 行工具栏最后一个选项(图标为一本书),即帮助(help)。然后按照“帮助(help)→SAS 产品→Base SAS→SAS 窗口引用→SAS 窗口索引”的步骤显示全部 SAS 窗口列表。

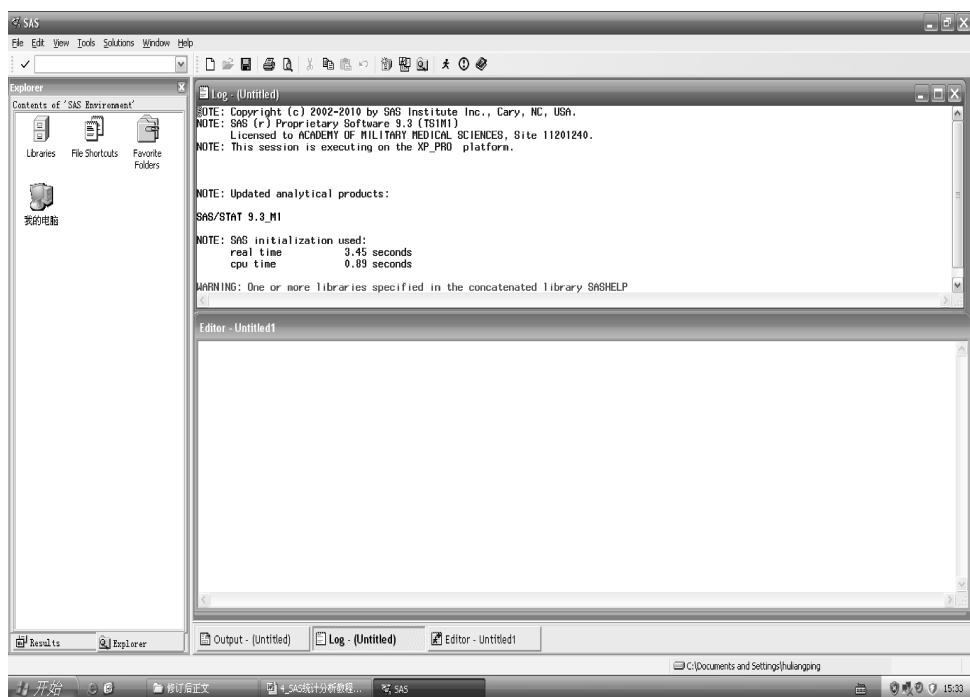


图 1-1 进入 SAS 9.3 后的主界面

1.1.3 SAS 过程与 SAS 程序

SAS 过程是 SAS 软件中经过编译后的程序，这些程序解决问题所依赖的理论和方法是公认的，因此，可以做到标准化、程序化和系统化。然而，用户要解决的问题却是千变万化的，对于用户的数据是什么，存放在何处，都是事先无法预知的。用户在调用某个具体的 SAS 过程之前，必须将上述信息传递给 SAS 系统，这些信息必须依据 SAS 语言规则来组织，它们被称为 SAS 引导程序，简称为 SAS 程序。

1.1.4 运行 SAS 软件的两种常用方式

在 SAS 系统中运行 SAS 软件通常有两种方式。第一种是非编程法（或称为菜单驱动法），即用户不需要编写 SAS 程序就可以直接调用 SAS 过程实现希望达到的目的。事实上，当用户通过菜单驱动系统选择某些项时，SAS 系统内部就在进行自动编程（即自动产生 SAS 程序），当用户的选择工作结束时，SAS 程序也就全部被产生出来，故也实现了用户的目的。但并非所有的任务都能通过此法来实现。第二种是编程法，即用户亲自在“程序编辑”窗口中写 SAS 程序（或直接调用别人事先写好的 SAS 程序）并提交给 SAS 系统执行。提交 SAS 程序的方法为按图 1-1 中上方第 2 行功能键倒数第 4 个图标（一个小人图像）。

本章仅介绍如何用编程法运行 SAS 软件。

1.1.5 SAS 程序结构

SAS 程序通常由两部分组成：一部分用于提供待分析的数据，称为 SAS 数据步；另一部分调用 SAS 系统中已编译过的能处理某个具体问题的真正程序，称为 SAS 过程步。一段 SAS 程

序可以有多个 SAS 数据步和多个 SAS 过程步,有时,也可以仅有其中的一种。若仅有 SAS 数据步,就需要利用 SAS 语言编写程序,以达到某个特定的研究目的;若仅有 SAS 过程步,必须事先提供可供调用的数据(称为 SAS 数据集)。先通过下面的一个简单实例,来直观了解这些抽象的概念。

【例 1-1】 两总体率差异性检验问题。从两个同类和规模相近的大工厂里随机地各抽取 254 个相同规格的零件,其中甲厂出现的次品数为 5 个;乙厂出现的次品数为 10 个。请问:甲、乙两工厂生产的这种零件的次品率之间的差别是否具有统计学意义?

【分析与解答】 这是两样本率比较问题,更确切地说,是关于两个总体率是否相等的假设检验问题,通常称为四格表资料的统计分析。可通过在“SAS 编程”窗口中输入下面的 SAS 程序来实现,设程序名为 SASTJFX1_1.SAS。

```
DATA rate;
    DO a=1 TO 2; DO b=1 TO 2;
        INPUT f @@ ; OUTPUT;
    END; END;
CARDS;
    5 249
    10 244
;
RUN;
ODS HTML;
PROC FREQ DATA=rate;
    WEIGHT f;
    TABLES a*b / CHISQ;
RUN;
ODS HTML CLOSE;
```

程序说明与修改指导: (1) SAS 数据步。从“DATA rate;”语句开始到第一个“RUN;”语句结束的这一段 SAS 程序被称为 SAS 数据步,其功能是创建 SAS 数据集,即为 SAS 系统提供待分析的数据所需要的 SAS 程序,包括与数据有直接联系的变量名。这里,变量 a 代表工厂类别,即代表两行的总名称,a=1 代表第一行上的甲工厂、a=2 代表第二行上的乙工厂;变量 b 代表产品检查结果,即代表两列的总名称,b=1 代表次品、b=2 代表合格品;f 代表与特定工厂及特定检查结果对应的产品数,即四格表资料中每个格子上的频数,如 249 代表来自甲工厂的合格品。

(2) SAS 过程步。介于“ODS HTML;”与“ODS HTML CLOSE;”两语句(其作用是使 SAS 输出结果以网页格式呈现)之间的语句段被称为过程步,即从“PROC FREQ DATA=rate;”到最后的“RUN;”语句的这段 SAS 程序被称为 SAS 过程步,其功能是调用 SAS 软件中能达到用户要求的真正程序(即 FREQ 过程)并产生相应的输出结果。

1.1.6 简单 SAS 程序中的 SAS 语句简介

全部 SAS 语句内容十分丰富,因篇幅所限,这里仅介绍 SAS 程序 SASTJFX1_1.SAS 中所涉及的几个 SAS 语句的大致含义。

每个 SAS 语句以分号(注意:必须是英文状态下的分号)结束,一行上可写多个 SAS 语句;不区分字母的大小写。上节这段 SAS 程序的数据步涉及 8 种 SAS 语句,现扼要介绍如下。

(1) DATA 语句,创建一个名为 rate 的临时 SAS 数据集(一旦退出 SAS 系统,它就消失了)。

(2)DO 语句,循环体的起始语句。

DO 语句与 END 语句成对出现,构成一个循环体。在上节 SAS 程序中,第一个 DO 语句与第二个 END 语句构成一个外循环,第二个 DO 语句与第一个 END 语句构成一个内循环。循环体的功能是创建标识性变量,同时使循环体内部的语句被重复执行规定的次数。

两个 DO 语句中的 a 和 b 被称为循环变量,DO 语句中 TO 前面的数被称为循环变量将取的最小值(或起始值),TO 后面的数被称为循环变量将取的最大值(或终止值)。这段 SAS 程序中外循环变量执行两次,内循环变量先后各执行两次,故对循环体内的两个语句而言,总共被执行了 4 次。

(3)INPUT 语句,创建一个名为 f 的变量,通过它去读取 CARDS 语句和空语句(即独占一行的分号)之间的数据,该语句中的“@@”为指针控制符,当 INPUT 语句中变量的个数少于一行上数据个数时,必须加上这两个符号,以确保完整地读入全部数据。

若 INPUT 语句中包含字符型变量,例如,若数据中性别变量 SEX 的具体取值为 SEX = MALE 代表男性、SEX = FEMALE 代表女性,则在 INPUT 语句中,变量 SEX 应写成“SEX \$”。若数据中第 1 列为人的姓名,其中最长的姓名占 18 个字符,不仅要 NAME 写成“NAME \$”,还应在 INPUT 语句前加一个定义字符型变量长度的语句,即“LENGTH NAME \$ 18. ;”。

(4)OUTPUT 语句,即输出语句,将 INPUT 语句变量 f 读取的每一数值送到循环体外面去(即放入计算机缓冲区中),以免后读取的数据将先读取的数据覆盖掉。

(5)END 语句,循环体结束语句。

(6)CARDS 语句,标志着其后为数据。

(7)空语句,以一个“;”独占一行的语句被称为空语句,标志着数据行的结束。千万不要将这个分号与最后一个数据写在同一行上。

(8)RUN 语句,SAS 数据步中的该语句标志着 SAS 数据步的结束。此语句可以不写,但写上显得完整。

上节这段 SAS 程序的过程步涉及 4 种 SAS 语句,现扼要介绍如下。

(1)PROC 语句,它是 SAS 过程步开始的标志,其后跟随着一个具体的 SAS 过程名,本例中为 FREQ,它是用于分析定性资料的一个 SAS 过程,其后写“DATA = rate”,它是一个选择项,意味着待分析的 SAS 数据集名为 rate。若在此之前程序中只有一个数据集,可以不加此选择项。

(2)WEIGHT 语句,指明 SAS 数据集中哪个变量代表列联表资料中各格上的频数,本例为变量 f。

(3)TABLES 语句,指明 SAS 数据集中哪个定性变量为行变量(本例为 a)、哪个为列变量(本例为 b),“/”代表其后为选择项,CHISQ 选项要求 FREQ 过程对 SAS 数据集中的数据进行 χ^2 检验。

(4)RUN 语句,SAS 过程步中的该语句标志着 SAS 过程步的结束,此语句不可省略。

初学者只需将自己的四格表资料中的 4 个数正确地替换掉此段程序中的 4 个数,将 SAS 程序提交给 SAS 系统执行(按 SAS 窗口上方第二行菜单条倒数第 4 个按钮,即小人像)即可。

何为正确替换?就本例而言,来自一个工厂的两个数据必须写在同一行(或列)上,而同一种检查结果必须写在同一列(或行)上。还应注意:检查结果必须按“次品与合格品”频数来表达,不能是“次品数与被检查总数”。

1.1.7 SAS 语言简介

1. SAS 语言概述

在编写像 SASTJFX1_1.SAS 那样的 SAS 程序时,所涉及的全部内容可概括为 SAS 语言。以笔者之见, SAS 语言由基本 SAS 语言(如 SAS 文件、基本 SAS 语句、常用 SAS 函数、SAS 数组和 SAS 过程等)和高级 SAS 语言(如宏、SAS SQL、SAS ODS、SAS/IML 等)两部分组成。这些内容十分丰富,需要大量篇幅才能对它们做一粗略的介绍。详细内容请查阅 SAS 说明书或从 SAS 软件的帮助窗口中查询,下同,不再赘述。

2. SAS 函数概述

SAS 中提供的常用 SAS 函数近 500 个,如求 -12.3 的绝对值,其对应的 SAS 语句为:“ $Y = \text{ABS}(-12.3);$ ”,也可这样写:“ $X = -12.3; Y = \text{ABS}(X);$ ”;再比如,求 49 的平方根,其对应的 SAS 语句为:“ $Y = \text{SQRT}(49);$ ”,也可这样写:“ $X = 49; Y = \text{SQRT}(X);$ ”。有些函数使用起来就没有这么简单了,因为那些函数中可能带有多个参数,需要搞清楚这些参数的含义,才能正确调用那些函数,获得用户需要的结果。

3. SAS 过程概述

SAS 软件中有 30 多个模块,每个模块中一般都有几十个 SAS 过程。这些 SAS 过程都能完成些什么任务,如何正确调用这些 SAS 过程,也不是轻而易举能交代清楚的。像 SASTJFX1_1.SAS 中所调用的 SAS 过程名为“FREQ”,总共写了 4 个 SAS 语句即可完成,其实,仅这一个 SAS 过程的全部语句和选择项所涉及的内容就相当多,不花上大约一周时间,可能无法全部掌握它。

1.1.8 SAS 数据集简介

1. SAS 数据集的概念

无论是采用非编程法还是编程法运行 SAS,首先都必须创建 SAS 数据集。直接在 SAS 编辑窗口录入原始数据,将其存储在某个外部设备上,只能称其为一个外部数据文件,不能称其为 SAS 数据集。何为 SAS 数据集呢?数据只有经过 SAS 系统加工后并按特定方式存储才能被称为 SAS 数据集,它将变量名及其取值(即具体数据)有机地结合在一起。

2. SAS 数据集的种类

SAS 数据集分为两类:一类被称为临时 SAS 数据集(如程序 SASTJFX1_1.SAS 中的数据集 rate),它被存储在 SAS/WORK 库(即文件夹)中,一旦退出 SAS 系统,它就消失了;另一类被称为永久 SAS 数据集,即使退出 SAS 系统,它仍被保留。这种数据集必须被保存在非 SAS/WORK 库中,可以是 SAS 系统默认的某个库,也可以是用户自己创建的某个库(即文件夹)。

3. 创建 SAS 数据集的方法

创建临时 SAS 数据集很简单,只需要像程序 SASTJFX1_1.SAS 那样去做就可以了。下面介绍一种创建永久 SAS 数据集的方法。

【例 1-2】 创建永久 SAS 数据集的方法。假定在 D 盘上有一个文件夹名为 SASTJFX,试将程序 SASTJFX1_1.SAS 中的数据步改造成能创建永久 SAS 数据集,并将永久数据集存储在 D 盘 SASTJFX 文件夹中。

【分析与解答】 所需要的 SAS 数据步如下,设程序名为 SASTJFX1_2.SAS。

```
LIBNAME table 'D:\SASTJFX';
DATA table.rate;
    DO a=1 TO 2; DO b=1 TO 2;
        INPUT f @@ ; OUTPUT;
    END; END;
CARDS;
5 249
10 244
;
RUN;
```

程序说明: LIBNAME 语句是创建文件关联名的重要语句,其后引号中的内容为路径(包括盘符和文件夹名),table 是用户自己取的关联名(一般不要超过 8 个字符),它就代表其后所写的路径。在 DATA 语句中,写“table.rate”,就意味着要创建一个名为 rate 的永久 SAS 数据集,它被存储在 D 盘 SASTJFX 文件夹中,存储后的实际数据集名为:rate.sas7bdat。

4. SAS 数据集名的种类

通常, SAS 数据集名有两类:一类被称为“一词名”,另一类被称为“两词名”。如上所述,临时 SAS 数据集为一词名(如 rate),而永久 SAS 数据集为两词名(如 table.rate)。在“一词名”中又可细分为以下三种。

(1) 用户自己取的一个词(如 rate)。

(2) 直接将 DATA 语句写成“DATA;”的形式,用户未给数据集取名, SAS 系统会自动给数据集取名,当执行一次 SAS 数据步后,系统将所创建的 SAS 数据集取名为 DATA1,再运行一次,取名为 DATA2……

(3) 使用 SAS 系统保留的特殊数据集名,如 _DATA_、_NULL_ 和 _LAST_。可以写成“DATA _DATA_;”,与写成“DATA;”是等价的,将给各次创建的数据集依次命名为 DATA1, DATA2, …; 若写成“DATA _NULL_;”则表明 SAS 系统将执行数据步,可用“PUT”等 SAS 语句输出中间结果,但观测值并不被写入 SAS 数据集 _NULL_,这样可以节省计算机资源;用“DATA _LAST_;”时,表明在此之前(指自此次进入 SAS 系统以来)无论创建了多少个 SAS 数据集,只调用最后被创建的那个 SAS 数据集。

1.1.9 如何利用 SAS 帮助窗口

与 SAS 语言、SAS 过程和 SAS 用法对应的内容非常丰富,谁也记不全那么多内容。怎么办?只要用户知道一些基本的线索,就可以通过 SAS 软件提供的丰富帮助功能来查询。通过单击 SAS 窗口界面上菜单栏第一行或第二行最后一个按钮,就可进入 SAS 帮助窗口,也可在 SAS 窗口界面左上角的命令盒内发送命令,如 HELP FREQ,回车后,就可直接查询有关 SAS 中 FREQ 过程的全部信息。

1.2 SAS 用法简介

1.2.1 初学者学习 SAS 的快捷方式

上节中,在介绍 SAS 程序结构和常用 SAS 语句时已顺便介绍了如何用编程法使用 SAS。由此可见,使用 SAS 并不是一件非常困难的事。难就难在 SAS 语言(包括 SAS 语句、SAS 函数、SAS 过程、SAS 高级编程技术)内容很多,需要花很多时间去学习和实践。

笔者给初学者提供一种学习 SAS 快速入门的方法，即调用他人编制好的 SAS 程序，只要解决了“对号入座”问题（即每个程序是干什么的），用自己的数据替换掉已有的 SAS 程序中的数据，将程序发送给 SAS 系统去执行，就可获得自己所需要的计算结果。每次结合他人的 SAS 程序和程序语句的讲解，一次学一点，不需多长时间，自己就慢慢掌握了很多常用的 SAS 语言，也就是说，边学边解决实际问题，不仅不会望而生畏，而且见效很快。

当数据少时，直接将数据写在程序中即可，但是当数据量很大时，这样做就不够方便了，尤其是遇到第三方格式数据，SAS 是不能直接读取的。如何用 SAS 进行实验设计、进行资料表达等内容，虽然很简单，但却是必须了解的内容，下面就这些基本内容做一扼要介绍。

1.2.2 实际运行 SAS

什么叫实际运行 SAS？若拟采用非编程法运行 SAS，只需根据 SAS 说明书中所交代的步骤去“点菜单”就能获得所需要的结果，这是第一种实际运行 SAS 的方法。第二种实际运行 SAS 的方法：在 SAS 程序编辑器中输入一段正确的 SAS 程序，如将本章第 1.1 节介绍的 SAS 程序发送给 SAS 系统执行，就称为实际运行 SAS，会产生如下的输出结果：

| FREQ 过程 | | | | |
|---------|---------|-------|-------|--------|
| 频数 | a * b 表 | | | |
| 百分比 | a | b | 合计 | |
| 行百分比 | 1 | 2 | | |
| 列百分比 | 1 | 5 | 249 | 254 |
| | | 0.98 | 49.02 | 50.00 |
| | | 1.97 | 98.03 | |
| | | 33.33 | 50.51 | |
| | 2 | 10 | 244 | 254 |
| | | 1.97 | 48.03 | 50.00 |
| | | 3.94 | 96.06 | |
| | | 66.67 | 49.49 | |
| | 合计 | 15 | 493 | 508 |
| | | 2.95 | 97.05 | 100.00 |

这是对输入的原始数据的详细描述，每个格内第一行为观察的频数；第二行为百分比，即以每个格上的频数为分子，以总频数为分母计算得到的相对数；第三行为行百分比；第四行为列百分比。

| a * b 表的统计量 | | | |
|--------------------|-----|---------|--------|
| 统计量 | 自由度 | 值 | 概率 |
| 卡方 | 1 | 1.7174 | 0.1900 |
| 似然比卡方 | 1 | 1.7497 | 0.1859 |
| 连续校正卡方 | 1 | 1.0991 | 0.2945 |
| Mantel-Haenszel 卡方 | 1 | 1.7140 | 0.1905 |
| Phi 系数 | | -0.0581 | |
| 列联系数 | | 0.0580 | |
| Cramer 的 V | | -0.0581 | |

这是用四种方法分析四格表资料所得到的结果。第一种为一般 χ^2 检验, $\chi^2 = 1.7174$, $P = 0.1900$; 第二种为似然比 χ^2 检验, $\chi^2 = 1.7497$, $P = 0.1859$; 其他从略。

最后三行为度量列联表中行变量与列变量间关联性强弱的系数, 其绝对值越接近于 1, 表明关联越密切。因未对其进行假设检验, 故无太大参考价值。

| Fisher 精确检验 | |
|------------------|--------|
| 单元格 (1,1) 频数 (F) | 5 |
| 左侧 Pr <= F | 0.1472 |
| 右侧 Pr >= F | 0.9435 |
| 表概率 (P) | 0.0907 |
| 双侧 Pr <= P | 0.2944 |

样本大小 = 508

以上是采用 Fisher 精确法分析四格表资料所得到的结果。

统计和专业结论: 因 $\chi^2 = 1.7174$, $P = 0.1900$, 说明两个工厂该产品的次品率之间的差别无统计学意义。虽然乙厂的次品数是甲厂次品数的 2 倍, 但 2 个次品率 (即 $P_{\text{甲}} = 5/254 = 1.97\%$ 与 $P_{\text{乙}} = 10/254 = 3.94\%$) 之间的差别无统计学意义。说明 2 个工厂此种零件的次品率接近相等, 即质量水平相当。

1.2.3 从实验设计角度谈 SAS 用法

与实验设计有关的内容可大致分为三类: 其一是进行随机化 (SAS 中有 PLAN 过程等); 其二是估计样本含量和检验效能 (SAS 中有 POWER 过程和 GLMPOWER 过程等); 其三是给出实验设计方案 (具体地说, 就是与特定设计类型对应的可用于安排实验的设计表格, SAS 中有多 个过程可用于此目的)。以上内容都可以通过非编程法和编程法来实现, 具体做法参见本书有关章节。

1.2.4 从资料录入角度谈 SAS 用法

1. 按数据库格式录入统计资料

人们收集的科研资料往往错综复杂, 但绝大部分统计资料都可表达成如表 1-1 所示的形式, 它常被称为“数据库格式”的复合型统计资料。这种呈现资料的方式把每个变量在每个个体身上的具体取值都清楚地展示出来了, 可以说是比较准确的原始资料。

【特例】 如何用数据库格式呈现多因素多指标资料。有人对 103 例冠心病患者 ($G = 1$) 和 100 例正常对照者 ($G = 2$) 进行了多项指标的观测, 资料见表 1-1。请问: 如何在 SAS 系统中录入这些资料, 以便于采用 SAS 软件对数据进行各种统计表达与描述或进行各种统计分析?

表 1-1 冠心病人与正常人多项指标的观测结果

| 编号 | 组别 | 性别 | 年龄 | 高血压史 | 吸烟史 | 胆固醇 | 甘油三酯 | 低密度脂蛋白 | 高密度脂蛋白 | 脂蛋白 α | 载脂蛋白 As | 载脂蛋白 B | 基因型 xbaI | 基因型 EcoRI | 服药情况 |
|-----|-----|-------|-------|-------|-------|-------|-------|--------|--------|--------------|----------|----------|----------|-----------|-------------|
| N | G | X_1 | X_2 | X_3 | X_4 | X_5 | X_6 | X_7 | X_8 | X_9 | X_{10} | X_{11} | X_{12} | X_{13} | X_{14} |
| 1 | 1 | 男 | 60 | 无 | 无 | 223 | 205 | 122 | 30 | 106 | 0.92 | 0.74 | -/- | -/- | 未服 |
| 2 | 1 | 女 | 46 | 无 | 无 | 166 | 51 | 84 | 57 | 56 | 1.14 | 0.54 | -/- | +/- | β 阻滞剂 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 203 | 2 | 男 | 69 | 有 | 无 | 224 | 110 | 58 | 49 | 132 | 1.10 | 0.96 | -/- | +/+ | 未服 |

若这些数据是写在纸上的,则只能在 SAS 软件的编辑窗口内一行一行地输入数据。每行代表一位受试者的全部信息(在 SAS 中,称其为一个观测),通常,同一行上的数据之间用空格符隔开;每一列代表一个变量在不同受试者身上的具体取值。变量的含义很广,它可以代表第 1 列的序号,可以代表第 2 列的组别,……,可以代表最后一列的合并用药情况。变量代表的内容可以是辅助信息,可以是分组标志或影响因素,可以是定量或定性的观测结果。

①若在 SAS 编辑窗口内输入的仅仅是数据(包括字符型数据),将其以文本格式存入外部媒介(硬盘、优盘或软盘)上,就被称为“数据文件”。这种格式的数据文件可以在 SAS 编辑窗口中编写 SAS 程序实现调用,调用的关键 SAS 语句是 INFILE 语句和 INPUT 语句。

例如,在 SAS 程序编辑窗口中输入例 1-1 的数据,形式如下:

```
1 1 5
1 2 249
2 1 10
2 2 244
```

说明:第一列代表工厂编号,“1”代表甲厂、“2”代表乙厂;第二列代表产品检查结果,“1”代表次品、“2”代表正品;第三列代表各条件下的样品数。按下面的方法可将此数据存储在 D 盘 SASTJFX 文件夹内,形成数据文件,假定数据文件名为 PRODUCT.DAT:

在 SAS 程序编辑窗口左上角:文件(FILE)→Save As(另存为)→在弹出的“另存为”窗口的左上角寻找并确定路径:D/SASTJFX→在此窗口下方倒数第二行的文件名命令盒内输入数据文件名:PRODUCT→在此窗口倒数第一行的文件类型命令盒中选择“.DAT”作为文件的扩展名→单击“保存(或“确定”)”按钮。

【例 1-3】如何将文本格式的数据读入 SAS 编程窗口。如何将数据文件读入 SAS 编程窗口,进行四格表资料的各种统计分析呢?在 SASTJFX1_1.SAS 程序中,只需将数据步修改成如下形式,而不需改动过程步。

【分析与解答】所需的 SAS 数据步如下,设程序名为 SASTJFX1_3.SAS。

```
DATA rate;
    INFILE 'D:\SASTJFX\product.dat';
    INPUT a b f;
ODS HTML;
PROC FREQ DATA=rate;
    WEIGHT f;
    TABLES a*b / CHISQ;
RUN;
ODS HTML CLOSE;
```

将这段 SAS 程序发送给 SAS 系统执行,可达到同样的效果。这种使用 SAS 的方法适合数据量很大的场合。

②若在 SAS 编辑窗口内输入数据,第一行输入了变量名,从第二行开始是变量的具体取值,也将其存成数据文件,这是错误的数据文件,不能用前述的方法被 SAS 系统直接调用,只能在日后给用户提个醒,每列数据的变量名是什么。若变量名写得比较科学,其含义一看便知,则对理解这些数据能起到“备忘录”的作用,否则,没有任何价值!

③若从 SAS 窗口通过“工具→表编辑器”方式进入表编辑器窗口,在此窗口内直接输入数据,窗口第一行带有变量名(用户可修改变量名),然后,将数据存入外部设备或计算机缓存区,就成了能被 SAS 系统直接调用的 SAS 数据集了。存入 SAS 系统自动创建的逻辑库

WORK 中的数据集会称为临时 SAS 数据集，存入其他位置的 SAS 数据集被称为永久 SAS 数据集。

2. 按实验设计类型录入统计资料

人们收集完数据后，有时，习惯将它们分类整理成一张统计表，统计表的分组标志通常是定性变量，而结果变量通常是定量的。这样的数据，当属于单因素设计定量资料时，常需进行 t 检验、单因素设计定量资料的方差分析或秩和检验；当属于某种多因素设计定量资料时，常需进行相应设计定量资料的方差分析。请看下面的例子。

【例 1-4】 如何使用多个 DO - END 循环语句读取多因素析因设计一元定量数据。某实验同时涉及 A、B、C 三个地位平等的实验因素，A 分为 2 个水平、B 分为 3 个水平、C 分为 4 个水平，观测指标为 OD 值，受试对象为样品，不同实验条件下均独立地重复做了 2 个样品，资料见表 1-2。请在 SAS 编辑窗口中输入此表中的主要变量和相应的数据，以便能进行相应设计定量资料的方差分析。

表 1-2 3 个实验因素作用下 OD 值的测定结果

| 因素 | | OD 值 | | | | | | | |
|----------------|----------------|-------|----------------|----------------|----------------|----------------|------|------|------|
| A 与 B | | 因素 C: | C ₁ | C ₂ | C ₃ | C ₄ | | | |
| A ₁ | B ₁ | | 0.39 | 0.41 | 0.37 | 0.39 | 0.42 | 0.38 | 0.44 |
| | B ₂ | | 0.37 | 0.36 | 0.43 | 0.45 | 0.41 | 0.37 | 0.42 |
| | B ₃ | | 0.45 | 0.43 | 0.46 | 0.39 | 0.38 | 0.35 | 0.39 |
| A ₂ | B ₁ | | 0.36 | 0.41 | 0.45 | 0.36 | 0.41 | 0.45 | 0.41 |
| | B ₂ | | 0.42 | 0.37 | 0.38 | 0.41 | 0.38 | 0.36 | 0.43 |
| | B ₃ | | 0.37 | 0.43 | 0.36 | 0.39 | 0.43 | 0.42 | 0.35 |

【分析与解答】 如果采取上例的方法输入数据，需要在每个定量数据前输入 4 个变量的水平代码。例如第一个数据 0.39，应当输入如下信息：1 1 1 1 0.39，这 4 个 1 分别代表因素 A、B、C 和重复实验次序都取 1 水平；同理，第二个数据 0.41，应当输入如下信息：1 1 1 2 0.41，…；最后一个数据 0.37，应当输入如下信息：2 3 4 2 0.37。这 4 个数分别代表因素 A 取 2 水平、因素 B 取 3 水平、因素 C 取 4 水平，而重复实验次序为第 2 次。显然，这样做太麻烦了，而且，很容易出错。简便的做法是，用 SAS 语言中的 DO - END 循环语句来自动产生因素 A、B、C 和重复实验次序的水平，其数据步如下，设 SAS 程序名为 SASTJFX1_4.SAS。

```
data doxunhuan;
  do A=1 to 2; do B=1 to 3;
    do C=1 to 4; do cixu=1 to 2;
      input OD @@ ; output;
    end; end; end; end;
  cards;
0.39 0.41 0.37 0.39 0.42 0.38 0.44 0.41
0.37 0.36 0.43 0.45 0.41 0.37 0.42 0.39
0.45 0.43 0.46 0.39 0.38 0.35 0.39 0.37
0.36 0.41 0.45 0.36 0.41 0.45 0.41 0.46
0.42 0.37 0.38 0.41 0.38 0.36 0.43 0.38
0.37 0.43 0.36 0.39 0.43 0.42 0.35 0.37
;
run;
```

程序说明：最外层的 DO - END 循环控制表中横向上水平变化最慢的变量(因素 A)，第二

层 DO-END 循环控制表中横向上水平变化较快的变量(因素 B), 这两个变量已将全部 6 行打上 A 与 B 的水平标记, 即前 3 行 A 均标记为 1、后 3 行 A 均标记为 2; 而 B 的标记从上到下分别为 1、2、3、1、2、3。每行有 8 个数据, 先按因素 C 分为 4 组, 其标记分别为 1、2、3、4, 每组内再按次序(cixu)分为标记 1、2。其效果是所形成的 SAS 数据集中的排列顺序与前面用“笨方法”产生的结果一致, 即(因篇幅太大, 中间部分省略了):

| | | | | |
|-----|-----|-----|------|------|
| A | B | C | cixu | OD |
| 1 | 1 | 1 | 1 | 0.39 |
| 1 | 1 | 1 | 2 | 0.41 |
| ... | ... | ... | ... | ... |
| 2 | 3 | 4 | 1 | 0.35 |
| 2 | 3 | 4 | 2 | 0.37 |

3. 按列联表类型录入统计资料

与前面的定量资料类似, 人们在表达多因素影响下的定性资料时, 习惯上将数据整理成列联表的形式, 特别是高维列联表资料, 很少用“数据库”的形式呈现资料, 见下面的例子。

【例 1-5】 如何使用多个 DO-END 循环语句读取多因素设计一元定性数据。某临床医生收集到如表 1-3 所示的资料, 请在 SAS 编辑窗口中输入此表中的主要变量和相应的数据, 以便能进行相应设计定性资料的统计分析。

表 1-3 甲、乙两种治疗方法对不同病程和不同病情某病患者的治疗效果

| 治疗方法 | 病程 | 病情 | 患者例数 | | | | |
|------|----|----|------|----|----|----|----|
| | | | 疗效: | 治愈 | 显效 | 好转 | 无效 |
| 甲 | 短 | 轻 | | 50 | 46 | 37 | 12 |
| | | 重 | | 42 | 35 | 32 | 23 |
| | 长 | 轻 | | 37 | 30 | 28 | 14 |
| | | 重 | | 31 | 24 | 25 | 38 |
| 乙 | 短 | 轻 | | 45 | 49 | 44 | 16 |
| | | 重 | | 38 | 43 | 39 | 22 |
| | 长 | 轻 | | 29 | 38 | 34 | 19 |
| | | 重 | | 22 | 33 | 30 | 28 |

注: 这个例子是假设的。

【分析与解答】 与前例相同, 输入数据的方法也有两种。第一种是在每个频数前需要提供 4 个变量的标记, 它们分别是治疗方法(treatment)、病程(time)、病情(degree)、疗效(effect)。这是很麻烦的事! 用 DO-END 循环语句就可方便地实现上述目标, 其 SAS 数据步如下, 设 SAS 程序名为 SASTJFX1_5.SAS。

```
data doxunhuan;
  do treatment = 'JIA', 'YI'; do time = 'short', 'long';
    do degree = 'light', 'weight';
      do effect = 'zhiyu', 'xianxiao', 'haozhuang', 'wuxiao';
        input number @@ ; output;
      end; end; end; end;
  cards;
50 46 37 12
42 35 32 23
37 30 28 14
31 24 25 38
```

```

45 49 44 16
38 43 39 22
29 38 34 19
22 33 30 28
;
run;

```

程序说明：最外层的 DO - END 循环控制表中横向上水平变化最慢的变量(治疗方法 treatment), 第二层 DO - END 循环控制表中横向上水平变化较快的变量(病程 time), 第三层 DO - END 循环控制表中横向上水平变化最快的变量(病情 degree)。这三个变量已将全部 8 行打上 treatment(治疗方法)、time(病程)、degree(病情)的水平标记, 即前 4 行 treatment 均标记为 JIA(甲), 后 4 行 treatment 均标记为 YI(乙); time 的标记从上到下分别为 short(短)、short(短)、long(长)、long(长)、short(短)、short(短)、long(长)、long(长); 而 degree 的标记从上到下依次是 light(轻)、weight(重)交替出现; 每行上的 4 列是 effect(疗效)的水平标记, 依次是 zhiyu(治愈)、xianxiao(显效)、haozhuang(好转)、wuxiao(无效), INPUT 语句中的 number 读取各行上的频数。形成的 SAS 数据集的样式与前例相似, 此处从略。

1.2.5 从不同格式数据转换角度谈 SAS 用法

1. 由非统计软件创建的数据文件与 SAS 数据集之间的互相转换

若待分析的数据已采用某些第三方非统计软件(如 Excel 软件等)创建了不同格式的数据文件, 其中有些可用 SAS 系统提供的导入数据接口方便地转换为 SAS 数据集(利用导出数据接口可以实现相反的操作)。请看下面的例子。

【例 1-6】 如何将一个 Excel 数据文件转变成 SAS 数据集。设有一个用 Excel 软件创建的数据文件 zhanghongleidata1.xls, 该文件中有两列数据, 第一列变量名为 A, 第二列变量为 B, A、B 的具体取值分别是计算机导航辅助方法与 CT 方法测定每一位骨病患者置入颈椎椎弓根螺钉的相对角度的数据。共有 140 对数据, 假定它们测自 140 位患者, 试将其转换为 SAS 数据集。

【分析与解答】 假定用 Excel 软件创建的数据文件 zhanghongleidata1.xls 存放在 D:\SAS-TJFX 内, 则通过如下步骤, 可将其转换为临时 SAS 数据集。

①进入 SAS 系统→文件→导入数据→在窗口右边弹出一个含有命令盒的窗口。

在命令盒中显示可导入的数据文件类型为: 97、2000 或 2002 年版的 Excel 软件产生的数据文件, 用户可通过此命令盒最右边的三角调整拟导入的数据文件的格式。

②单击窗口下边的“Next”按钮→弹出一个小窗口, 要求通过浏览方式确定拟导入的 Excel 文件的路径和文件名→选中 D:\SASTJFX\zhanghongleidata1.xls→单击“OK”按钮。

③系统询问要导入的 Excel 文件是表几(自动显示表 1, 即 sheet1\$, 若不是表 1, 可重新选择)→单击“Next”按钮。

④弹出一个新窗口, 有两个命令盒, 上行为逻辑库名(自动显示临时库名 WORK, 也可改变), 下行为拟创建的数据集名→输入 dao_hang_and_CT_data→单击“Finish”按钮。

⑤在窗口左边逻辑库中 WORK 库内就有刚创建的 SAS 数据集。

⑥用鼠标左键双击此数据集, 可显示此数据集的内容。

将 Excel 文件显示的界面与已转换后得到的 SAS 数据集做一个直观比较。

不难发现: 变量名被 SAS 系统修改了, A、B 分别被改为 F1 和 F2; Excel 文件中的第一行

数据被 SAS 系统“吃掉了”！造成这种不良后果的主要原因是在将 Excel 文件转换成 SAS 数据集的最后一步未取消系统中一个隐含的“设置”，在上面导入数据的第二步之后，会弹出一个窗口。其内的命令盒内有“Sheet1 \$”。在“Sheet1 \$”之下有一个“Options”按钮，单击此按钮会弹出另一个窗口。

在此窗口上的第一行是处于被选中的状态，即使用 Excel 文件中的第一行作为 SAS 数据集的变量名。若用户在 Excel 文件中第一行输入的是数据而不是变量名，则应将第一行中复选框内的“√”去掉。去掉后单击右边的“OK”按钮，接下来的操作步骤与上面的第④步到第⑥步相同，可获得正确的转换结果。

本例还可以通过在 SAS 编程窗口中运行下面的一段 SAS 程序，实现将 Excel 文件转成 SAS 数据集的目的。设程序名为 SASTJFX1_6.SAS。

```
PROC IMPORT OUT = WORK.ZHANGHONGLEI DBMS = EXCEL REPLACE
      DATAFILE = "D:\SASTJFX\zhanghongleidata1.xls";
      SHEET = "Sheet1 $ ";
      GETNAMES = YES;
RUN;
```

值得注意的是，若 Excel 文件中第一行不是变量名而是数据，则上面的程序倒数第二句应改为“GETNAMES = NO”；若要转换的数据在 Excel 文件的第三张表单中，则上面的程序倒数第三句应改为“SHEET = "Sheet3 \$"”。若转换成功，则新产生的临时 SAS 数据集 ZHANGHONGLEI 存放在 SAS/WORK 库中，可通过 SAS 资源管理器找到此库，双击此库中的 SAS 数据集名，便可将其打开，也可在编程窗口中直接调用这个临时 SAS 数据集。

2. 用 SAS 系统读入其他版本或分析软件创建的数据集

若待分析的数据已采用第三方统计软件(如 SPSS、BMDP 等统计软件包)创建了不同格式的数据集，则需要通过使用 libname 语句和在 SAS 中内置的转换程序(称为读取特定格式数据的库引擎)将特定的数据文件转换为 SAS 数据集。这种方式使用起来不很方便，下面介绍如何利用 SPSS 软件提供的文件存储功能将 SPSS 数据集转换成 SAS 数据集。

如果用户正在使用的计算机上正确地安装了 SPSS 软件，直接用鼠标左键双击 SPSS 数据文件进入 SPSS 系统并打开该文件，选择“另存为”，在弹出的存储文件的窗口内选择合适的“保存类型”并输入拟创建的 SAS 数据集名，确定后，就得到转换后的 SAS 数据集。

1.2.6 从资料表达角度谈 SAS 用法

SAS 中有些过程给出的结果并不太理想，如用 FREQ 过程生成定量资料的频数分布表，只要两个数据不完全相等，就形成两个分组标志，这样形成的频数分布表很长，显得过细，没有实用价值。实际上，可以在使用 FREQ 过程的基础上配合使用 FORMAT 过程，生成比较有实用价值的频数分布表(设程序名为 SASTJFX1_6.SAS)；还可按用户的需要去选定第一组的组段下限、组距、组数等要求，利用丰富的 SAS 语言编程，产生用户自己量身定制的频数分布表。

1.2.7 从统计分析角度谈 SAS 用法

若用户需要进行的统计分析在 SAS 软件中已有相应的过程(如单因素 2 水平设计定量资料 t 检验，有 TTEST 过程；两因素析因设计定量资料一元方差分析，有 ANOVA 过程和 GLM 过

程等；对列联表资料的分析，有 FREQ 过程和 CATMOD 过程等），此时，直接调用 SAS 过程实现统计分析就比较简单了。然而，对于某些计算问题，SAS 中尚无现成的 SAS 过程，此时，可以利用 SAS 语言并按已知的计算公式或算法编写 SAS 程序，从而实现统计分析。

总之，要想利用 SAS 解决自己的各种问题，除了要会调用 SAS 中的全部过程外，还应该全面掌握 SAS 语言。运用 SAS 语言，可以编写出当前已有但 SAS 过程解决得不满意的问题或尚解决不了的一些问题。如果用户使用某种其他编程语言（如 C 语言、C++ 语言或 Java 语言），需要写大量的代码来设置环境、构造窗口等，而用 SAS 语言，只需把精力放在要解决的具体问题上，编程工作量就小多了。此时，用户不难体会到，SAS 的用途的确是很广的。

1.3 本章小结

本章用很短的篇幅介绍了 SAS 软件和 SAS 用法。显然，此举旨在起到抛砖引玉之目的。因为 SAS 软件内容十分丰富，用 SAS 能解决的问题更是不胜枚举。本章从 SAS 软件结构、SAS 界面简介、SAS 过程与 SAS 程序、运行 SAS 软件的两种常用方式、SAS 程序结构、简单 SAS 程序中的 SAS 语句简介、SAS 语言简介、SAS 数据集简介和如何利用 SAS 帮助窗口这 9 个方面宏观地扼要介绍了 SAS 软件；接着，又从初学者学习 SAS 的快捷方式、实际运行 SAS、从实验设计角度谈 SAS 用法、从资料录入角度谈 SAS 用法、从不同格式数据转换角度谈 SAS 用法、从资料表达角度谈 SAS 用法和从统计分析角度谈 SAS 用法这 7 个方面概述了 SAS 用法。

（胡良平 李子建 刘惠刚）

第2章 SAS 语言基础介绍

由第1章内容可知，运行 SAS 的方法很多，通常分为两大类：其一，非编程法或称为菜单驱动法；其二，编程法。SAS 软件包已经有很多现成的程序，为何还要用户编程呢？事实上，SAS 软件包中被编译后的程序被称为“SAS 过程”，而用户编写的程序是为了向 SAS 系统提供数据和任务请求的，被称为 SAS 引导程序，简称 SAS 程序。本章将介绍编写 SAS 程序所需要的 SAS 语言基础。

2.1 SAS 数据步中常用 SAS 语句

SAS 程序(SAS PROGRAM)是 SAS 用户运用 SAS 语言编写的一段程序，其目的是为了将用户的实验数据与指标(即变量)名称联系在一起，并告诉 SAS 系统调用特定的 SAS 过程完成某项任务。SAS 程序通常可分为两部分，分别被称为 SAS 数据步(DATA STEP)和 SAS 过程步(PROC STEP)。请看下面的一个具体例子。

【例 2-1】 表 2-1 中是一段 SAS 程序，通过这段程序可初步了解 SAS 程序的结构和书写格式。设程序名为 SASTJFX2_1.sas。

表 2-1 8 名中学生的身高体重资料代码

| 行号 | 语 句 | 行号 | 语 句 |
|----|--|----|---|
| 1 | / *****/ | 20 | ods html; |
| 2 | 18 名中学生的身高体重资料 l | 21 | proc means data = SASTJFX2_1; |
| 3 | *****/ | 22 | var bmi; |
| 4 | options nodate number = 0; | 23 | output out = result mean = BMImean; |
| 5 | data SASTJFX2_1; | 24 | run; |
| 6 | input name \$ sex \$ age weight height @@; | 25 | proc print data = result(keep = BMImean); |
| 7 | BMI = weight/(height/100) ** 2; | 26 | format BMImean 4.1; |
| 8 | label BMI = 'body mass index'; | 27 | run; |
| 9 | datalines; | 28 | ods html close; |
| 10 | WANG M 14 42 150 | | |
| 11 | ZHANG M 16 46 170 | | |
| 12 | LI F 15 44 149 | | |
| 13 | TANG M 15 38 162 | | |
| 14 | LIU F 14 47 162 | | |
| 15 | DIAO F 16 52 168 | | |
| 16 | ZHU M 14 45 158 | | |
| 17 | JIA F 16 50 167 | | |
| 18 | ; | | |
| 19 | run; | | |

数据步包括要求 SAS 创建一个或几个新的 SAS 数据集的语句和对创建数据集所必须进行的运算操作语句。每个 DATA 步以 DATA 语句开头，可以包含任意多个 SAS 程序语句。编写报表、文件管理、信息检索等都在 DATA 步中予以实现。

2.1.1 数据获取语句

数据获取通常使用 input 语句。

【例 2-2】 用 input 语句以简单方式读取数据流中的数据。设程序名为 SASTJFX2_2.sas。

```
data SASTJFX2_2;
    input name $ weight height;
cards;
WANG 42 150
ZHANG 46 170
;
run;
proc print;
run;
```

上面的程序中，input 语句描述输入数据记录值的形式，给相应的变量赋值。\$ 在 name 之后，说明 name 是字符型变量，而不是数值型变量；而 weight 和 height 都是数值型变量。

输出结果：

| Obs | Name | Weight | Height |
|-----|-------|--------|--------|
| 1 | WANG | 42 | 150 |
| 2 | ZHANG | 46 | 170 |

【例 2-3】 如何用 input 语句以列方式读取数据。设程序名为 SASTJFX2_3.sas。

```
data SASTJFX2_3;
    input id 1 name $ 3-7 weight 9-11;
    list;
datalines;
1 WANG 42
2 ZHANG 46
;
run;
proc print;
run;
```

在 input 语句中，列数跟在变量名之后，指示输入数据记录中的变量值从哪些列中读取，并将读取值赋予相应的变量。本例从数据文件的第 1 列读取变量 id 的值，从第 3~7 列读取字符型变量 name 的值，从第 9~11 列读取变量 weight 的值。

数据行或外部文件数据排列方式如下：

```
RULE:  ---- + ----1---- + ----2---- + ...
      1 WANG 42
      2 ZHANG 46
```

读取的数据样式如下：

| id | name | weight |
|----|-------|--------|
| 1 | WANG | 42 |
| 2 | ZHANG | 46 |

【例 2-4】 如何用 input 语句以格式化方式读取数据。设程序名为 SASTJFX2_4.sas。


```
data SASTJFX2_4;
    input name $char5. +2 height comma6.;
datalines;
WANG HT150
;
run;
```

在 input 语句中，变量名后跟随变量的输入格式，由输入格式给出变量值类型、变量值宽度，并将读取的值赋予相应的变量。本例以 \$char5. 格式读取变量 name 的值，跳过 2 列后以 comma6. 格式读取变量 height 的值。

读取的数据样式如下：

| | name | height |
|---|------|--------|
| 1 | WANG | 150 |

【例 2-5】 如何用 input 语句以列表方式读取数据。设程序名为 SASTJFX2_5. sas。

```
data SASTJFX2_5;
    input name: $13. age;
datalines;
Wangli 14
Zhanghaizheng 16
;
run;
```

在 input 语句中，以简单方式列出变量名，SAS 扫描输入数据值，这种方式可以读取排成一列或由空格等分隔符分隔的数据值。冒号“:”允许 input 语句使用用户指定的一个输入格式读取变量值。对于字符型变量值，这种方式输入指针从一个非空白列开始读取变量值，直到下一个空白列，或到变量定义的长度，或者到达数据行的末端。当分隔符是空格时，本例使用 input 语句对变量长度不一致的变量规定统一长度（13 个字节）。

【例 2-6】 如何用 input 语句以命名方式读取数据。设程序名为 SASTJFX2_6. sas。

```
data SASTJFX2_6;
    input name = $ age =;
datalines;
name = WANG age = 14
name = ZHANG age = 16
;
run;
```

在 input 语句中，变量名后跟随一个等号“=”，SAS 读取变量名及等号之后的数据值，并将读取值赋予相应的变量。上面程序运行后，SAS 读取数据行中“name =”后的值赋给变量 name，读取“age =”后的值赋给变量 age。

读取的数据样式如下：

| | name | age |
|---|-------|-----|
| 1 | WANG | 14 |
| 2 | ZHANG | 16 |

【例 2-7】 如何用 input 语句以命名方式读取包含空格的字符型数据。设程序名为 SASTJFX2_7. sas。

```
data SASTJFX2_7;
  input header = $15. name = $;
  datalines;
  header = age =16 and up name = ZHANG
  ;
run;
```

数据行中的“age = 60 and up”是一个含有等号、空格的数据值，所以需要在其前后加两个空格，以区分其他数据值。

读取的数据样式为：

| header | name |
|-----------------|-------|
| age = 16 and up | ZHANG |

比较：

(1) input 语句可以读取外部文件或数据行中的数据值，并且需要对这些数据值的类型、长度等信息进行描述；而 set 语句则是读取 SAS 数据集中的数据，这些数据已经具有描述信息。

(2) input 语句是读取数据值，而 put 语句则是在日志窗口或外部文件中写入数据值或文本。

(3) input 语句能从外部文件中读取数据值，而 infile 语句是指定外部文件位置，并具有控制外部文件怎样读取的选项。

2.1.2 数据步文件管理语句

1. data 语句

该语句指示数据步开始，或为输出数据集提供名字。

【例 2-8】 如何用 data 语句规定要创建的 SAS 数据集。

```
data; /* 系统自动规定数据集名 datan, 第几次运行, n 就取几 */
data fitness; /* 创建临时数据集 fitness */
data _null_; /* 特殊名, 不创建 SAS 数据集, 用于输出 */
```

【例 2-9】 数据集选项举例

```
data new(drop = y); /* 去除 SAS 数据集 new 中的变量 y */
data new(label = '治疗方法'); /* 规定数据集 new 标签名为“治疗方法” */
```

【例 2-10】 如何用 data 语句创建 DATA 步数据视窗文件。

```
libname out 'c:\sas\mydir1';
data out.scores /view = out.scores;
```

在斜线后面跟选项 view =，该选项规定用户要创建的输入 DATA 步视图的名字，并且在 DATA 语句中必须规定两次这个名字。libname 语句是创建文件关联名的重要语句，其后引号中的内容为路径（包括盘符和文件夹名），out 是用户自己取的关联名（不要超过 8 个字符），它就代表其后所写的路径。在 DATA 语句中，写“out.scores”意味着要创建一个名为 scores 的永久 SAS 数据集，它被存储在 C 盘 SAS\mydir1 文件夹中，存储后的实际数据集名为：scores.sas7bdat。

【例 2-11】 如何用 data 语句存储被编辑程序或执行一个被存储的编辑程序。

```
libname stored 'c:\sas\mydir2';
data out.Stales2 /pgm=stored.scales;
    set sales 1;
    ...
run;
```

在斜线后面跟着选项 `pgm =`，并给出程序名称。`out` 同样为用户自己取的关联名，代表其后所写的路径。“`out.Stales2`”代表创建一个名为 `Stales2` 的永久 SAS 数据集，它被存储在 `sas\mydir2` 文件夹中，存储后的实际数据级名为 `Stales2.sas7bdat`。

SAS 数据集分为两类：一类称为临时 SAS 数据集，如例 2-8 中的数据集，它们被存储在 SAS/WORK 库（即文件夹）中，一旦推出 SAS 系统，它们就自动消失了；另一类被称为永久 SAS 数据集，即使退出 SAS 系统，它仍被保留，如例 2-10、例 2-11。这种数据集必须被保存在非 SAS/WORK 库中，可以是 SAS 系统默认的某个库，也可以是用户自己创建的某个库（即文件夹）。

2. cards 语句

该语句指示数据行开始。

【例 2-12】 如何用 `cards` 语句指示数据行开始。设程序名为 `SASTJFX2_12.sas`。

```
data SASTJFX2_12;
input x y;
    cards;
    [数据行]
    ;
run;
```

`cards` 语句等价于 `datalines` 语句，均可指示数据行开始。

3. cards4 语句

该语句指示包含分号的数据行开始。

【例 2-13】 如何用 `cards4` 语句读取含有分号的数据。设程序名为 `SASTJFX2_13.sas`。

```
data SASTJFX2_13;
    input name $ age;
    cards4;
    WANG; 14
    ZHANG; 16
    LI; 15
    ;;;
run;
```

`cards4` 语句等价于 `datalines4` 语句，均可指示包含分号的数据行开始。

4. file 语句

该语句用于规定将要输出的外部文件，它一般要与 `put` 语句配合使用。同一个 DATA 步可以用到多个 `file` 语句。`file` 语句是可执行语句，因而可以用在 (IF-THEN) 过程中。

【例 2-14】 如何用 `file` 语句在日志文件中显示 `hello` 字符串。设程序名为 `SASTJFX2_14.sas`。

```
data SASTJFX2_14;
    file log;
    put 'hello';
run;
```

【例 2-15】 如何用 file 语句在结果输出窗口的指定行列显示文本。设程序名为 SAS-TJFX2_15.sas。

```
data SASTJFX2_15;
    file print linesleft=remain pagesize=20;
    put @ 5 'name' @ 10 'sex' //
    @ 5 'age' @ 10 'height';
run;
```

@5 将指针移动到第 5 列, @10 将指针移动到第 10 列, /将指针前移到下一个输入行的第 1 列。

显示的文本如下:

```
name    sex
age      height
```

5. infile 语句

该语句用来定义一个外部数据文件, 文件中的数据用 input 语句来读取。外部文件可以是已存在的磁盘文件, 也可以是从键盘上输入的数据行。

【例 2-16】 如何用 infile 语句读取外部文件。设程序名为 SASTJFX2_16.sas。

```
filename mydata 'd:\SASTJFX\2_1.txt';
data SASTJFX2_16;
    infile mydata;
    input name $ sex $ age weight height;
run;
```

上面程序先用 filename 语句规定了 mydata 是操作系统相应目录下文本文件 2_1.txt 的文件引用名, 然后在 infile 语句中使用, 这种引用称为间接引用; 当然也可以直接引用文件的物理地址, 则程序中不需要 filename 语句, infile 语句改成如下所示:

```
infile 'd:\SASTJFX\2_1.txt';
```

【例 2-17】 如何用 infile 和 datalines 语句读取以逗号为分隔符的数据流数据。设程序名为 SASTJFX2_17.sas。

```
data SASTJFX2_17;
    infile datalines delimiter=',';
    input name $ age;
    datalines;
    WANG,14
    ZHANG,16
    LI,15
    ;
run;
```

上面程序中, 在 infile 语句中添加了 delimiter = ',' 选项, 通过 infile 和 datalines 语句的联合使用, 读取以逗号为分隔符的数据流数据。

【例 2-18】 如何用 infile 语句更新外部文件。设程序名为 SASTJFX2_18.sas。

```
data SASTJFX2_18;
  infile 'd:\SASTJFX\2_1.txt';
  file 'd:\SASTJFX\2_1.txt';
  input name $ sex $ age weight height;
  if name = 'WANG' then age = 17;
  put name $ sex $ age weight height;
run;
```

可以将 infile 语句和 file 语句联合使用更新外部文件，在上面的程序中，更新操作系统相应目录下的 2_1.txt 文件，如果文件中有 name = 'WANG' 的观测，则将这个观测的 age 变量的值改为 17。

【例 2-19】 如何用 infile 语句读取有连续分隔符的数据行数据。设程序名为 SASTJFX2_19.sas。

```
data SASTJFX2_19;
  infile cards dsd;
  input x @@ ;
  cards;
  ,5,,6
  ;
run;
```

由于使用了 dsd 选项，下面程序读取的数据值为缺失值、5、缺失值、6。如果没有 dsd 选项，添加 dlm = ',' 选项时，则只能读取 5 和 6 两个数据值。

比较：

infile 语句为 input 语句指定输入文件，file 语句为 put 语句指定输出文件。infile 语句通常用于确定外部文件，而 datalines 语句指示数据来自随后的数据流，但用户可以在 infile 语句中使用 datalines 选项来控制数据的读取。

6. put 语句

该语句在 SAS 日志、输出窗口或最近的 file 语句指定的外部位置写入语句。

【例 2-20】 如何用 put 语句在外部文件中写入变量值和文本字符串。设程序名为 SASTJFX2_20.sas。

```
data SASTJFX2_20;
  set SASTJFX2_1;
  file 'd:\SASTJFX\bmi.txt';
  put name $1-10 bmi 4.1;
run;
```

上面的程序将 SASTJFX2_20 数据集中的变量 name 和 bmi 写入操作系统相应目录下的 bmi.txt 文件。

【例 2-21】 如何用 put 语句输出行。设程序名为 SASTJFX2_21.sas。

```
data SASTJFX2_21;
  input name $ age;
  file print;
  put @ 5 name +10 10* '_' / @ 6 age /;
  datalines;
  WANG 14
```

```
ZHANG 16
LI 15
;
run;
```

上面的程序中，file print 语句指定 put 语句的输出位置在结果输出窗口，在 put 语句中，指定从第 5 列开始写入变量 name 的值，后移 10 列显示 10 个下画线，然后在下一行的第 6 列写入变量 age 的值，换行写下下一个内容。

【例 2-22】 如何用 put 语句以列方式写变量值。设程序名为 SASTJFX2_22.sas。

```
data SASTJFX2_22;
  x=11;
  y=15;
  put x 10 -18 .1 y 20 -28 .1;
run;
```

上面的程序中，指定在日志第 10~18 列以 1 位小数的形式写变量 x 的值，在日志第 20~28 列以 1 位小数的形式写变量 y 的值。

【例 2-23】 如何用 put 语句以格式化方式写变量值。设程序名为 SASTJFX2_23.sas。

```
data SASTJFX2_23;
  input name $ age;
  put @ 10 (name age) ($16. 8.1);
datalines;
WANG 14
ZHANG 16
LI 15
;
run;
```

上面的程序中，在 SAS 日志的第 10 列以 \$16. 格式写变量 name 的值，以 8.1 格式写变量 age 的值。

读取的数据样式如下：

| | |
|-------|------|
| WANG | 14.0 |
| ZHANG | 16.0 |
| LI | 15.0 |

【例 2-24】 如何用 put 语句以列表方式写变量值。设程序名为 SASTJFX2_24.sas。

```
data SASTJFX2_24;
  input name $ age;
  put name $ age ;
datalines;
WANG 14
ZHANG 16
LI 15
;
run;
```

上面的程序中，指定在日志中写变量 name、age 的值。

【例 2-25】 如何用 put 语句以命名方式写变量值。设程序名为 SASTJFX2_25.sas。

```

data SASTJFX2_25;
  input name $ age;
  put name = $ 6 -16 age =18 -21;
datalines;
WANG 14
ZHANG 16
LI 15
;
run;

```

上面的程序中，变量后跟一个等号，还具有输出格式和写入列的位置。

读取的数据样式如下：

| | |
|--------------|----------|
| name = WANG | age = 14 |
| name = ZHANG | age = 16 |
| name = LI | age = 15 |

比较：

(1) put 语句是在 SAS 日志或外部文件中写入变量值和文本字符串，而 input 语句是从输入数据流的数据行或外部文件读取原始数据。

(2) put 语句和 input 语句都使用单尾符@ 和双尾符@@ 在输出或输入缓存中保持当前行。不同的是，在 input 语句中，双尾符@@ 是在数据步从一次循环到下次循环时在输入缓存中保持行；而在 put 语句中，单尾符@ 和双尾符@@ 具有相同的作用，都是在数据步循环中保持行。

(3) put 语句和 output 语句都能在数据步产生输出。put 语句使用输出缓存，并将输出行写入外部位置、SAS 日志或用户的显示器中；而 output 语句使用程序数据向量，并将观测写入 SAS 数据集。

7. modify 语句

该语句在一个已存在的数据集的某位置替换、删除和追加观测，不产生另外的副本。

【例 2-26】 如何用 modify 语句修改所有观测。设程序名为 SASTJFX2_26.sas。

```

data SASTJFX2_26;
  modify SASTJFX2_1;
  BMI = sqrt(BMI);
run;

```

上面的程序中，将 BCJQ3_1 数据集中的变量 BMI 开方。

8. merge 语句

该语句将两个或多个 SAS 数据集的观测合并成一个观测。

【例 2-27】 如何用 merge 语句以一对一的方式合并多个数据集。设程序名为 SASTJFX2_27.sas。

| | |
|--|--|
| <pre> data SASTJFX2_27A; input num name \$ sex \$; cards; 1 WANG M 2 ZHANG M 3 LI F </pre> | <pre> data SASTJFX2_27; merge SASTJFX2_27A SASTJFX2_27B; proc print; run; </pre> |
|--|--|

```
4 TANG M
;
data SASTJFX2_27B;
input t1-t3;
cards;
89 76 90
78 88 74
96 98 92
;
run;
```

合并后的数据集为：

| Obs | num | name | sex | t1 | t2 | t3 |
|-----|-----|-------|-----|----|----|----|
| 1 | 1 | WANG | M | 89 | 76 | 90 |
| 2 | 2 | ZHANG | M | 78 | 88 | 74 |
| 3 | 3 | LI | F | 96 | 98 | 92 |
| 4 | 4 | TANG | M | . | . | . |

【例 2-28】 如何用 merge 语句以变量 x 匹配合并多个数据集。设程序名为 SASTJFX2_28. sas。

```
data SASTJFX2_28A;
input num name $ sex $;
cards;
1 WANG M
2 ZHANG M
3 LI F
4 TANG M
;
data SASTJFX2_28B;
input num t1 - t3;
cards;
1 89 76 90
3 78 88 74
4 96 98 92
;
run;

proc sort data = SASTJFX2_28A;
by num; run;
proc sort data = SASTJFX2_28B;
by num; run;
proc print data = SASTJFX2_28A; run;
proc print data = SASTJFX2_28B; run;
data SASTJFX2_28;
merge SASTJFX2_28A SASTJFX2_28B;
by num; run;
proc print;
run;
```

合并后的数据集为：

| Obs | num | name | sex | t1 | t2 | t3 |
|-----|-----|-------|-----|----|----|----|
| 1 | 1 | WANG | M | 89 | 76 | 90 |
| 2 | 2 | ZHANG | M | . | . | . |
| 3 | 3 | LI | F | 78 | 88 | 74 |
| 4 | 4 | TANG | M | 96 | 98 | 92 |

比较：

- (1) merge 语句从两个或多个 SAS 数据集中合并观测；update 语句从两个数据集中合并观测；update 语句改变或更新主数据集中的观测值，也可以添加观测。
- (2) 像 update 语句一样，modify 语句也是通过改变或更新主数据集中的观测从两个数据集中合并观测。
- (3) 使用 set 语句读取两个或多个数据集中的观测，也有些类似于使用 by 语句的 merge 语句，但两者的功能不完全一样。

9. update 语句

该语句通过申请处理更新主文件。

【例 2-29】 如何用 update 语句更新数据集。设程序名为 SASTJFX2_29.sas。

```
data data1;
    input num name $ sex $;
    cards;
    1 WANG M
    2 ZHANG M
    3 LI F
    4 TANG M
    ;
data data2;
    input num name $;
    cards;
    1 WANG
    2WU
    3.
    4 TANG
    ;
run;
```

```
data SASTJFX2_29;
    update data1 data2;
    by num;
proc print;
run;
```

上面的程序用处理数据集 data2，以变量 num 的值更新主数据集 data1 中的数据。

更新后的数据集为：

| Obs | num | name | sex |
|-----|-----|------|-----|
| 1 | 1 | WANG | M |
| 2 | 2WU | M | |
| 3 | 3 | LI | F |
| 4 | 4 | TANG | M |

比较：

- (1) update 语句和 merge 语句都能更新数据集中的观测。
- (2) merge 语句自动用第 2 个数据集中的缺失值取代第 1 个数据集中已存在的值；但 update 语句不会默认这样做，除非指定 updatemode = nomissingcheck 选项。
- (3) update 语句申请处理改变或更新主文件中被选择观测的值，update 语句不能添加新观测。

10. set 语句

该语句从一个或多个 SAS 数据集中读取观测。

【例 2-30】 如何用 set 语句保留部分观测。设程序名为 SASTJFX2_30.sas。

```
data SASTJFX2_30 (drop = sex age);
    set work.SASTJFX2_1;
    if sex = 'M';
run;
```

上面的程序从 SASTJFX2_30 数据集中复制男生的部分数据。

新生成的数据集如下：

| name | weight | height | BMI |
|-------|--------|--------|---------|
| WANG | 42 | 150 | 18.6667 |
| ZHANG | 46 | 170 | 15.9170 |
| TANG | 38 | 162 | 14.4795 |
| ZHU | 45 | 158 | 18.0260 |

【例 2-31】 如何用 set 语句连接数据集。设程序名为 SASTJFX2_31.sas。

```
data SASTJFX2_31A;
    input a b @@ ;
cards;
1 1 2 3 3 5 4 7 5 9
data SASTJFX2_31B;
    input a c @@ ;
cards;
1 2 2 4 3 6 4 8 5 10
data SASTJFX2_31;
    set SASTJFX2_31A SASTJFX2_31B;
proc printdata = SASTJFX2_31;
run;
```

本例将 SASTJFX2_31A、SASTJFX2_31B 依次相连,形成有 a、b、c 三变量的 SASTJFX2_31 数据集。

| obs | a | b | c |
|-----|---|---|----|
| 1 | 1 | 1 | . |
| 2 | 2 | 3 | . |
| 3 | 3 | 5 | . |
| 4 | 4 | 7 | . |
| 5 | 5 | 9 | . |
| 6 | 1 | . | 2 |
| 7 | 2 | . | 4 |
| 8 | 3 | . | 6 |
| 9 | 4 | . | 8 |
| 10 | 5 | . | 10 |

假设上例中 SASTJFX2_31A 的变量 b 与 SASTJFX2_31B 中变量 c 实际是同一变量(意义相同),则要把它们拼接在一起,可如下处理:

```
set SASTJFX2_31A SASTJFX2_31B(rename = (c=b));
```

或

```
set SASTJFX2_31A(rename = (b=job)) SASTJFX2_31B(rename = (c=job));
```

输出结果如下:

| obs | a | b(job) |
|-----|---|--------|
| 1 | 1 | 1 |
| 2 | 2 | 3 |
| 3 | 3 | 5 |
| 4 | 4 | 7 |

| | | |
|----|---|----|
| 5 | 5 | 9 |
| 6 | 1 | 2 |
| 7 | 2 | 4 |
| 8 | 3 | 6 |
| 9 | 4 | 8 |
| 10 | 5 | 10 |

比较：

(1) set 语句从一个已存在的 SAS 数据集中读取观测；input 语句是从外部文件或数据流中读取数据，用来建立 SAS 变量和为变量赋值。

(2) 在 set 语句中使用“key =”选项可以不按顺序、通过数据值来读取观测；在 set 语句中使用“point =”选项可以不按顺序、通过观测号读取观测。

(3) set 语句可以将两个或更多的数据集连在一起，形成一个单独的大的数据集，属“纵向连接”。merge 语句将两个或多个 SAS 数据集中的观测值合并成一个新数据集中的单个的观测值，属“横向合并”。

11. by 语句

该语句控制数据步复制、合并、修改或更新等数据集操作，规定特殊的分组变量。

【例 2-32】 如何用 by 语句将数据集中的观测按变量 age 降序进行排列。

```
by descending age;
```

在数据步中，by 语句只用于 set、merge、modify 或 update 语句中。默认情况下，SAS 认为数据集按数值或字母升序排序，descending 选项对数据集按指定变量的降序排列。

2.1.3 SAS 变量操作语句

1. array 语句

该语句用于定义数组元素。

【例 2-33】 如何用 array 语句定义一维数组。

```
array simple{3} red green yellow;
```

用 array 语句定义了一个一维数组，名为 simple，数组中有 red、yellow、blue 3 个元素。

```
array test{6} t1 -t6 (10,20,2* (40 50));
```

用 array 语句定义了一个名为 test 的一维数组，数组中 6 个元素名分别为 t1、t2、t3、t4、t5、t6，其初始值分别为 10、20、40、50、40、50。

```
array test[7] _temporary_(2* (1:3),0);
```

用 array 语句定义了一个名为 test 的一维临时数组，数组中 7 个元素的初始值分别为 1、2、3、1、2、3、0。

【例 2-34】 如何在赋值语句中引用数组。设程序名为 SASTJFX2_36.sas。

```
data SASTJFX2_36;
  array num{3} n1 -n3 (1,1,0);
  do i=1 to 3;
    if num{i}=0 then num{i}=1;
  end;
run;
```

程序 SASTJFX2_36 中，首先定义了一个 3 个元素的一维数组，数组元素的初始值分别为 1、1、0，然后使用 do 语句处理数组中的每个元素，在 if 选择结构语句中，使用赋值语句对符合条件的数组元素进行了重新赋值，即将初始值为 0 的元素值改为 1。

2. attrib 语句

该语句将一个或多个变量与其输入、输出格式、标签或者长度等属性联系在一起，即定义或修改变量的属性。

【例 2-35】 如何用 attrib 语句对单个变量定义多种属性。

```
attrib name length = $20 label = 'the name of students';
```

将字符型变量 name 的长度定义为 20 字节，标签为 'the name of students'。

【例 2-36】 如何用 attrib 语句对多个变量定义相同的属性。

```
attrib x1 x2 x3 length = $5;
```

将变量 x1、x2、x3 的长度定义为 5 字节。

3. length 语句

该语句指定储存变量的字节数。

【例 2-37】 如何用 length 语句指定储存变量的字节数。设程序名为 SASTJFX2_40.sas。

```
data SASTJFX2_40;
  length name $ 13;
  input name sex $ age weight height;
  cards;
  WangPing M 14 42 150
  ZhangYingJie M 16 46 170
  LiLai F 15 44 149
  ;
```

用 length 语句指定变量 name 为 13 字节。由于 name 变量已在 length 语句中定义为字符型变量，故 input 语句中可以不再用 \$ 号作定义。

4. label 语句

该语句定义变量的描述标签信息。

【例 2-38】 如何用 label 语句为变量设置标签。设程序名为 SASTJFX2_41.sas。

```
proc print data = SASTJFX2_41 noobs label;
  var name sex age weight height;
  label name = '姓名' sex = '性别' age = '年龄' weight = '体重' height = '身高';
run;
```

运行结果如下：

| SAS 系统 | | | | |
|--------|----|----|----|-----|
| 姓名 | 性别 | 年龄 | 体重 | 身高 |
| WANG | M | 14 | 42 | 150 |
| ZHANG | M | 16 | 46 | 170 |
| LI | F | 15 | 44 | 149 |
| TANG | M | 15 | 38 | 162 |
| LIU | F | 14 | 47 | 162 |

| | | | | |
|------|---|----|----|-----|
| DIAO | F | 16 | 52 | 168 |
| ZHU | M | 14 | 45 | 158 |
| JIA | F | 16 | 50 | 167 |

5. format 语句

该语句在 DATA 步中把变量与输出格式联系起来。输出格式可以是 SAS 提供的格式，也可以是用户使用 PROC FORMAT 定义的格式。当 SAS 系统打印这些变量的值时，将使用所联系的格式来输出这些值。DATA 步使用 format 语句可永久地把格式同变量联系起来。PROC 步用 format 语句指定的格式仅仅在 PROC 步起作用。

【例 2-39】 如何用 format 语句使变量以自定义格式输出。设程序名为 SASTJFX2_42.sas。

```
data SASTJFX2_42;
  input weight @@ ;
  format weight5.2;
  datalines;
  42 46 44
  ;
run;
proc print;
run;
```

上面的程序用 format 语句规定 weight 变量以 5.2 格式输出。

运行结果如下：

| Obs | weight |
|-----|--------|
| 1 | 42.00 |
| 2 | 46.00 |
| 3 | 44.00 |

比较：

format 和 attrib 语句都可以定义和修改变量的输出格式，format 语句可用于 DATASETS 过程，改变或取消变量的输出格式。

6. informat 语句

该语句将变量与其输入格式联系在一起，即定义变量的输入格式。

【例 2-40】 如何用 informat 语句规定临时的默认输入格式。设程序名为 SASTJFX2_43.sas。

```
data SASTJFX2_43;
  informat default = $ char4. default = 4.2;
  input name $ sex $ age weight height;
  put name $ sex $ age weight height;
  cards;
  WANG M 14 42 150
  ;
run;
```

程序提交后，LOG 窗口输出显示：

WANG M 0.14 0.42 1.5

例中，在 input 语句列出的变量没有规定输入格式，那么使用这里规定的默认输入格式，即用格式 \$ char4. 输入 name 和 sex，用格式 4.2 输入 age、weight、height。

7. drop 语句

该语句将变量排除在数据集之外。

【例 2-41】 如何用 drop 语句删除变量。设程序名为 SASTJFX2_44. sas。

```
data SASTJFX2_44;
    set SASTJFX2_1;
    drop sex;
proc print;
run;
```

上面的程序运行后删去了 SASTJFX2_1 中的变量 sex。

运行结果为：

| Obs | name | age | weight | height | BMI |
|-----|-------|-----|--------|--------|---------|
| 1 | WANG | 14 | 42 | 150 | 18.6667 |
| 2 | ZHANG | 16 | 46 | 170 | 15.9170 |
| 3 | LI | 15 | 44 | 149 | 19.8189 |
| 4 | TANG | 15 | 38 | 162 | 14.4795 |
| 5 | LIU | 14 | 47 | 162 | 17.9089 |
| 6 | DIAO | 16 | 52 | 168 | 18.4240 |
| 7 | ZHU | 14 | 45 | 158 | 18.0260 |
| 8 | JIA | 16 | 50 | 167 | 17.9282 |

8. keep 语句

该语句指定包含在 SAS 输出数据集中的变量。

【例 2-42】 如何用 keep 语句保留变量。设程序名为 SASTJFX2_45. sas。

```
data SASTJFX2_45;
    set SASTJFX2_1;
    keep name age;
proc print;
run;
```

上面的程序运行后仅保留 SASTJFX2_45 中的变量 name 和 age。

运行结果为：

| Obs | name | age |
|-----|-------|-----|
| 1 | WANG | 14 |
| 2 | ZHANG | 16 |
| 3 | LI | 15 |
| 4 | TANG | 15 |
| 5 | LIU | 14 |
| 6 | DIAO | 16 |
| 7 | ZHU | 14 |
| 8 | JIA | 16 |

比较：

(1) keep 语句不能用于过程步，而“keep =”选项可以用于过程步。

(2) keep 语句可应用于 data 语句命名的所有输出数据集，如果要在不同的数据集中写入不同的变量，可以使用“keep = ”数据集选项。

(3) drop 语句是与 keep 语句平行的语句，drop 语句用于指定从输出数据集中排除的变量。

(4) keep 和 drop 语句选择出包括或排除在数据集中的变量，子集 if 语句则是根据条件选择某些观测。

(5) 不要混淆 keep 和 retain 语句，retain 语句是使 SAS 从数据集一次循环至下一个循环时保持变量的值。

9. rename 语句

该语句为输出数据集中的变量指定新的名字。

【例 2-43】 如何用 rename 语句重命名变量。设程序名为 SASTJFX2_46. sas。

```
data SASTJFX2_46;
    input name $ age;
    rename name = sname;
    put name;
datalines;
WANG 14
ZHANG 16
LI 15
;
run;
```

上面的程序中，使用 rename 语句将变量 name 改变为 sname，可以看到在 put 语句中还是使用变量的旧名。

比较：

(1) rename 语句不能用于过程步，而“rename = ”选项可用于过程步。

(2) “rename = ”数据集选项允许用户指定在输入或输出数据集中想要重命名的变量，处理前在输入数据集中使用。

(3) 如果在输出数据集中使用“rename = ”数据集选项，当前数据步的程序中必须继续使用变量的旧名，当输出数据集产生后，可以使用变量的新名。

(4) set 语句中的“rename = ”选项在输入数据集中重命名变量，可以在当前数据步程序中使用新变量名。

(5) 重命名变量作为一项管理任务，可使用 DATASETS 过程或通过 SAS 窗口界面存取变量，这些方法更简单，而且不需要数据步处理。

10. retain 语句

该语句使 input 语句或赋值语句生成的变量从数据步一次循环到下次循环时保持它的值。可以使用 retain 语句为个别变量、变量列表或数据成员赋初始值。

【例 2-44】 如何用 retain 语句赋初始值。

```
retain month1 -month5 1 year0 a b c 'XYZ';
```

将变量 month1 ~ month5 的初始值设为 1，将变量 year 的初始值设为 0，将变量 a、b、c 的初始值设为 'XYZ'。

```
retain month1 -month5(1:4);
```

将 month1、month2、month3、month4 的初始值分别设为 1、2、3、4，变量 month5 未设初始值。

2.1.4 SAS 观测值操作语句

1. abort 语句

该语句停止执行当前数据步处理、SAS 正进行的作业或者会话。

【例 2-45】 如何用 abort 语句使 SAS 程序有计划地终止。设程序名为 SASTJFX2_48.sas。

```
data SASTJFX2_48;
    input theight @@ ;
        if height > 175 | height < 150 then abort;
datalines;
150 170 149 162 162 168 158 167
;
run;
```

abort 语句常出现在 if-then 或 select 这种分支结构语句中，可以规定出现错误的条件，使 SAS 程序有计划地终止。上面的程序提交运行后，当遇到 height 大于 175 或 height 小于 150 的数据值满足 if 语句条件时，在日志窗口显示了一条错误信息“abort 语句在某行某列终止了执行”和一条关于 SASTJFX2_48 数据集可能不完整的警告信息，产生的数据集中只包含前 2 个观测，第 3 个观测及其后的观测都停止了处理。

2. stop 语句

该语句可以停止执行当前数据步。

【例 2-46】 如何用 stop 语句停止执行当前数据步。

```
if idcode = 9999 then stop;
```

当变量 idcode 的值为 9999 时，用 stop 语句停止数据步执行。

比较：

stop 语句在窗体操作和交互行操作方式下，也可以停止当前处理，但 abort 语句会将自动变量“_error_”的值设置为 1，而 stop 语句则不会；在批处理和非交互行操作方式下，abort 语句会将自动变量“_error_”设置为 1，并终止处理，而 stop 语句是终止当前处理，继续进行后续语句的处理。

3. 赋值(分配)语句 (Assignment Statement)

该语句求得等号“=”右边表达式的值，并将结果值储存在等号“=”左边的变量中。

【例 2-47】 如何用赋值语句将变量 x 的原值加 1，然后赋予变量 x。

```
x = x + 1;
```

【例 2-48】 如何用赋值语句将字符串 Tom 赋予变量 name。

```
name = 'Tom';
```

【例 2-49】 如何用赋值语句将复合表达式的返回值赋予变量 y。

```
y = not0 - exp(n/(n-1)) + x + 3;
```


【例 2-50】 如何用赋值语句将返回值赋予变量 z。设程序名为 SASTJFX2_53.sas。

```
data SASTJFX2_53;
    length x $2.;
    input (x y) ($);
    z = x || y;
    datalines;
a b
;
run;
```

上面的程序中，首先定义了 x 是 2 字节长度的字符型变量，y 也为字符型变量，未定义长度，默认为 8 字节。用串联符(||)串联变量 x 和 y，通过赋值语句将返回值赋予变量 z，所以变量 z 也为字符型变量，长度为 10 字节。

4. 累加语句 (Sum Statement)

该语句将表达式的结果与一个累加变量相加。

【例 2-51】 如何用累加语句将变量 bal 与变量 deb 的相反数相加。

```
bal + (-deb);
```

注意这里的加号“+”是必须有的，否则不能构成累加语句。

【例 2-52】 如何用累加语句将变量 nx 与表达式“x ne .”的返回值相加(表达式成立则返回值为 1，不成立返回值为 0)。

```
nx + (x ne .);
```

【例 2-53】 在 if-then 条件结构语句中使用累加语句，如果符合 if 语句指定的条件则执行累加语句。

```
if status = 'ready' then OK + 1;
```

比较：

累加语句相当于联合使用 sum() 函数和 retain 语句：

```
retain 变量 0;
变量 = sum(变量, 表达式);
```

5. declare (dcl) 语句

该语句声明一个数据步的组件对象，产生一个实例并为数据步对象初始化值。

【例 2-54】 如何用 declare 语句产生一个数据步名为 h 的 hash 组件对象。

```
declare hash h();
```

上面的语句相当于下面的 2 个语句：

```
declare hash h;
h = _new_hash();
```

6. _new_ 语句

该语句产生数据步组件对象的实例。

【例 2-55】 如何用 _new_ 语句创建 hash 对象并分配给变量 h。

```
declare hash h;
h = _new_hash();
```

其中, declare 语句告诉 SAS 变量 h 是一个 hash 对象。

7. delete 语句

该语句停止处理当前观测。

【例 2-56】 如何用 delete 语句将变量 leafwt 中为缺失值的观测删除。

```
if leafwt = . then delete;
```

【例 2-57】 如何用 delete 语句和 if-then 语句使符合条件的观测不被写入数据集中。设程序名为 SASTJFX2_60.sas。

```
data SASTJFX2_60;
    input name $height;
    if height < 150 then delete;
datalines;
WANG 150
ZHANG 170
LI 149
TANG 162
LIU 162
DIAO 168
ZHU 158
JIA 167
;
run;
```

比较:

(1) 当指定一个条件从数据集中排除某些观测, 或遇到当前观测不需要继续处理时, 使用 delete 语句更简便; 当指定一个条件包含某些观测时, 使用 if 子集语句更简便。

(2) drop 语句用于排除某些变量, delete 语句用于排除某些观测。

(3) remove 语句也具有 delete 语句相似的功能和用法, 但 remove 语句只用于 modify 语句中, 而且可以是物理或逻辑删除某些观测。

8. remove 语句

该语句从 SAS 数据集中删除一个观测。

【例 2-58】 如何用 remove 语句删除 SASTJFX2_61 数据集中 name 为 LIU 的观测。设程序名为 SASTJFX2_61.sas。

| | |
|---|--|
| <pre>data SASTJFX2_61; input name \$height; datalines; WANG 150 ZHANG 170 LI 149 TANG 162 LIU 162 DIAO 168 ZHU 158 JIA 167 ; run;</pre> | <pre>data SASTJFX2_61; modify SASTJFX2_61; if name = 'LIU' then remove; run;</pre> |
|---|--|

比较:

- (1) 如果数据步有 output、replace 或 remove 语句，必须有明确地将观测输出的语句。
- (2) output、replace 和 remove 语句相互依赖，不止一个语句可以应用于同一个观测，它们的次序只是逻辑次序。
- (3) 如果 output、replace 和 remove 语句都应用于同一个观测，最后才执行 output 动作，以保持观测指针位置的正确。
- (4) remove 语句可执行物理或逻辑删除，可用于 modify 语句中所有的数据集引擎；delete 和子集 if 语句都是物理删除，所以它们在 modify 语句的一些引擎中无效。

9. replace 语句

该语句在相同位置替换一个观测。

【例 2-59】 根据一定条件，利用处理数据集 temp 中的数据，替换主数据集 SASTJFX2_62 中变量 age 的值。设程序名为 SASTJFX2_62.sas。

```
data SASTJFX2_62;
    input name $ age;
    datalines;
WANG 14
ZHANG 16
LI 15
;
run;
data temp;
    input name $ age;
    datalines;
WANG 19
ZHANG 21
LI 20
;
run;
```

```
data SASTJFX2_62;
    modify SASTJFX2_62 temp;
    by name;
    if age > 20 then replace;
run;
```

新生成的数据集如下：

| name | age |
|-------|-----|
| WANG | 14 |
| ZHANG | 21 |
| LI | 15 |

10. output 语句

该语句将当前观测写入 SAS 数据集。

【例 2-60】 在给定的条件满足时，如何用 output 语句将当前观测写入 SAS 数据集。

```
if deptcode gt 2000 then output;
```

【例 2-61】 如何用 output 语句将 age > 30 的观测写入 temp 数据集。

```
if age > 30 then output temp;
```

【例 2-62】 如何用 output 语句从一个输入文件创建多个数据集。设程序名为 SASTJFX2_65.sas。

```

data SASTJFX2_65 temp1 temp2;
  input name $ age;
  if 30 >= age > 25 then output SASTJFX2_65;
  if age > 30 then output temp1;
  else output temp2;
datalines;
WANG23
ZHANG31
LI28
;
run;

```

当遇到 $25 < \text{age} \leq 30$ 的观测时，将观测输入到 SASTJFX2_65 数据集中；当 $\text{age} > 30$ 时将观测写入 temp1 数据集，否则，将观测写入 temp2 数据集（注意 else 语句和最近的一个 if 语句相匹配，temp2 数据集中包括了 $\text{age} \leq 30$ 的所有观测）。

比较：

(1) output 语句将当前观测写入数据集；put 语句将变量值或文本写入外部文件或 SAS 日志。

(2) 控制一个观测写入指定的数据集，使用 output 语句；控制某些变量写入数据集，可以在 data 语句中使用“keep = ”、“drop = ”数据集选项，也可以使用 keep 或 drop 语句。

(3) 当在 modify 语句中使用 output 语句时，使用 output、replace 或 remove 语句取代数据步结尾默认的动作，如果在数据步使用这些语句，需要对加入数据集中的新观测有明确的输出语句。

output、replace 和 remove 语句相互独立，只要在顺序合理，不止一个语句可应用于同一个观测。如果 output 和 replace 或 remove 语句都作用于一个给定的观测，最后执行 output 语句，以保持观测指针位置的正确。

11. error 语句

该语句设置自动变量“_error_”的值为 1，有选择性地在日志中写入信息。

【例 2-63】 如果符合条件，如何用 error 语句在日志窗口显示信息和变量 age 值。

```

if type = 'teen' & age > 19 then
  error 'type and age did not match' age = ;

```

12. if 语句，子集

该语句的作用是当观测遇到表达式指定的条件时，则继续处理。

【例 2-64】 如何用 if 语句实现如果 age 不为缺失值或 0，则执行处理。

```

if age;

```

【例 2-65】 如何用 if 语句建立只包含 sex = 'F' 的观测的数据集。设程序名为 SASTJFX2_68.sas。

```

data SASTJFX2_68;
  input (name sex) ( $ ) @@ ;
  if sex = 'F';
datalines;
WANG M
ZHANG M
LI F
;
run;

```

比较：

(1) 子集 if 语句相当于下面的 if-then 语句：

```
if not(表达式) then delete;
```

(2) 当生成数据集时，如果指定一个包含观测的条件比较容易，则使用子集 if 语句；如果指定一个排除观测的条件比较容易，则使用 if-then、delete 语句。

13. call 语句

该语句调用 call 子程序。

【例 2-66】 如何用 call 语句将变量 x1、x2、x3 设为缺失值。

```
call missing(of x1 - x3);
```

【例 2-67】 如何用 call 语句产生声音。

```
call sound(523,2000);
```

产生一个频率为 523Hz(中央 C 音)、持续 2000 毫秒的声音。

【例 2-68】 如何用 call 语句发布操作系统命令。设程序名为 SASTJFX2_71.sas。

```
data SASTJFX2_71;
  call system('dir *.txt');
run;
```

调用 system() 函数向操作系统发布 'dir *.txt' 命令，列出操作系统工作目录下所有 txt 文件信息。

【例 2-69】 如何用 call 语句调用随机函数子程序产生随机数。设程序名为 SASTJFX2_72.sas。

```
data SASTJFX2_72;
  retain x110 x2 15;
  do i = 1 to 5;
    call rannor(x1,y1);
    call rannor(x2,y2);
    output;
  end;
run;
```

每个随机函数都有一个相应的 call 子程序，当一个数据步产生多组随机数时，随机函数产生的多组随机数字来自同一个数字流，而使用 call 子程序比使用随机函数能更好地控制种子值，产生的多组数字由于种子随机独立，所以各组随机数字之间也相互独立(参见第 4 章)。

14. list 语句

该语句将被处理的输入数据观测记录写入 SAS 日志。

【例 2-70】 如何用 list 语句在日志中显示变量长度记录的长度。设程序名为 SASTJFX2_73.sas。

```
data SASTJFX2_73;
  input name $ age;
  if age < 16 then list;
  datalines;
WANG 14
```

```
          ZHANG 16
          LI 15
;
run;
```

上面的程序运行后，日志中将显示一个标尺，并显示记录的长度。

```
RULE:      ---- + ----1---- + ----2---- + ----3 - ...
1          WANG 14
2          LI 15
```

比较：
list 和 put 语句的区别见表 2-2。

表 2-2 list 和 put 语句的区别

| 动作 | list 语句 | put 语句 |
|-----------|--------------------|--------------------------------|
| 何时写 | 在数据步每次循环结束 | 立即 |
| 写什么 | 输入数据记录 | 变量或文本 |
| 写在哪 | 只能写在 SAS 日志中 | 可以写在 SAS 日志、输出结果窗口或外部文件中 |
| 可以和哪些语句合用 | 只有 put 语句 | 任何数据读取的语句 |
| 处理十六进制数 | 如果遇到非显示字符自动显示十六进制数 | 只有给定了十六位进制的输出格式，才能以十六进制的形式显示字符 |

15. lostcard 语句

当 SAS 遇到数据中一个观测具有多条记录，记录中有缺失值或无效记录时，该语句校正输入数据。

【例 2-71】 下面的程序中，假定数据行第 2 个观测的第 2 个记录丢失，SAS 将第 3 个观测的第 1 个记录读成第 2 个观测的第 2 个记录，则可能出现与预想情况不符的问题。设程序名为 SASTJFX2_74.sas。

```
data SASTJFX2_74;
  input x y;
  datalines;
1 2
3
5 6
;
run;
```

上面程序生成的数据集样式如下：

| | | |
|-----|---|---|
| Obs | x | y |
| 1 | 1 | 2 |
| 2 | 3 | 5 |

在上面程序中加入 lostcard 语句运行后，数据集中没有观测，日志中显示如下信息：

```
NOTE: LOST CARD.
RULE:      ---- + ----1---- + ----6---- + ----3---- + ----4---- + ----5...
226      1 2
NOTE: LOST CARD.
```

228 5 6

NOTE: LOST CARD.

NOTE: LOST CARD.

229 ;

x = . y = . _ERROR_ = 1 _N_ = 1

NOTE: INPUT 语句到达一行的末尾, SAS 已转到新的一行。

NOTE: 数据集 WORK.BCJQ3_49 有 0 个观测和 2 个变量。

【例 2-72】 如何在条件结构语句中使用 `lostcard` 语句识别缺失数据记录并校正输入数据。设程序名为 `SASTJFX2_75.sas`。

```
data SASTJFX2_75;
  input id1 -3 age
        #2 id2 1 -3 loc
        #3 id3 1 -3 wt;
  if id ne id2 or id ne id3 then do;
    put 'DATA RECORD ERROR: ' id=id2=id3 =;
  lostcard;
end;
datalines;
301 32
301 61432
301 127
302 61
302 83171
400 46
409 23145
400 197
411 53
411 99551
411 139
;
run;
```

数据步读取数据时, 如果 `id` 不匹配(`id` 不等于 `id2`, 或 `id` 不等于 `id3`), 则 SAS 执行 `put` 语句和 `lostcard` 语句。本例中, `id` 为 302 的观测缺少一行记录(本应还有一行 `id` 为 302 的记录), `id` 为 400 的观测的第 2 行记录错误(`id` = 409, 本应为 400), 所以数据行中, 只有前 3 行和最后 3 行被当成 2 个观测写入数据集(本应有 4 个观测), 数据行中间的 5 行数据则执行了 `lostcard` 语句, 被丢弃, 所以生成的数据集中只包含数据行前 3 行和最后 3 行形成的 2 个观测。

16. 空语句(Null Staments)

该语句指示数据行结束, 或相当于一个占位符。

语法 ; 或 ;;;;

空语句不执行任何动作, 但它也是执行语句。`datalines` 或 `cards` 语句后的空语句主要用于指示数据行结束; 标号语句可以位于空语句之前, 空语句也可以用在条件处理中。

17. where 语句

该语句是在执行数据集连接(`SET`)、合并(`MERGE`)、更新(`UPDATE`)或修改(`MODIFY`)之前进行的操作。使用 `where` 语句时, 因为 SAS 系统只从输入数据集中读入满足条件的观测, 所以这样的 SAS 程序更有效。

【例 2-73】 如何用 where 语句从 SAS 数据集中选取相应的观测。设程序名为 SASTJFX2_76.sas。

```
data temp;
    input name $ age;
datalines;
WANG 14
ZHANG 16
LI 15
;run;
data SASTJFX2_76;
    set temp;
    where age < 15;
run;
```

上面的程序中，在 temp 数据集中读取 age < 15 的观测形成 SASTJFX2_76 数据集。

比较：

在 DATA 步 where 语句和子集 if 语句的最大差别是 where 语句在观测读入到程序数向量之前起作用，而子集 if 语句对已经在程序数据向量的观测起作用；where 语句不是执行语句，而子集 if 语句是可执行语句；where 语句有自己的表达式，而子集 if 语句使用“SAS 表达式”；where 语句仅仅从 SAS 数据集的观测中选择，而子集 if 语句可以从已存在的 SAS 数据集中或在用 input 语句产生的观测中选择；当 where 表达式和子集 if 表达式产生相同结果时，在几乎所有情况下，where 要比 if 的效率高；从大的 SAS 数据集中选择一个小的子集时用 where 语句比用子集 if 语句效率高得多，因为在进行选择之前 SAS 系统没有从大数据集中移动所有观测到这个数据集的程序数据向量中。

不要将 where 语句和 drop 或 keep 语句混淆。drop 和 keep 语句用于选择变量，where 语句用于选择观测。

18. missing 语句

该语句在输入数据中分配一个字符代表特殊的数字型缺失值。

【例 2-74】 如何用 missing 语句指定字符代表特殊的缺失值。设程序名为 SASTJFX2_77.sas。

```
data SASTJFX2_77;
    missing a r;
    input name $ age;
datalines;
WANG 14
ZHANG A
LI R
;run;
```

上面的数据中，用 a 代表被调查者当时不在家，用 r 代表被调查者拒绝回答问题，这种方式比使用无效的数据值更清楚。

比较：

“missing = ”系统选项允许用户指定一个字符显示在普通缺失值的地方。如果数据中包含代表特殊缺失值的字符，比如 a 或 z，不要使用“missing = ”系统选项来定义它们，应使用 missing 语句。

2.1.5 数据步循环与控制语句

1. do 语句

do 语句与 end 语句功能介绍见表 2-3。

表 2-3 do 语句与 end 语句功能介绍

| 语句 | 功能 | 语句 | 功能 |
|----------|-----------------------------|----------|-------------------------|
| do | 创建一组语句 | do while | 重复执行 do 循环中的语句至某个条件不再为真 |
| 循环 do | 根据下标变量重复执行 do 和 end 语句之间的语句 | do over | 在 do 循环中对隐含数组元素执行一些语句 |
| do until | 重复执行 do 循环中的语句至某个条件为真 | end | 标记一个 do 组或 select 组结束 |

【例 2-75】 如何在 IF/THEN 语句中用 do 语句。设程序名为 SASTJFX2_78. sas。

```
if years > 5 then do;
  months = years * 12;
end;
else yrsleft = 5 - years;
```

本例用简单 do 语句实现只有 years 大于 5 时，才执行 do 组语句，如果 years 小于或等于 5 则不执行 do 组语句，继续执行 else 语句后的赋值语句。

【例 2-76】 如何用 do 循环语句读取 2 × 2 列联表资料。设程序名为 SASTJFX2_79. sas。

```
data SASTJFX2_79;
  do i = 1 to 2;
    do j = 1 to 2;
      input f @@ ;
      output;
    end;
  end;
datalines;
30 10 11 49
;
run;
```

上面的程序使用套嵌 do 循环语句将具有 2 水平的两组受试对象的频数资料创建为 SAS 数据集。

【例 2-77】 如何用 do while 语句实现当 n 小于 5 时总是执行 do 循环。设程序名为 SAS-TJFX2_80. sas。

```
n = 0;
do while (n < 5);
  put n =;
  n + 1;
end;
```

【例 2-78】 如何用 do until 语句计算 π 的近似值。设程序名为 SASTJFX2_81. sas。

```
data SASTJFX2_81;
  retain n1 t1 pi 0 s1;
  do until (abs(1/n) < 1e-7));
    pi = pi + t;
    n = n + 2;
    s = -s;
    t = s/n;
  end;
```

```

output;
end;
pi=pi*4;
put pi;
run;

```

π 的近似值可由下面公式得到： $\pi/4 \approx 1/1 - 1/3 + 1/5 - 1/7 \cdots$ 直到最后一项的绝对值小于 10^{-7} 为止。程序运行后，put 语句在日志窗口显示 pi 值为 3.1415926536。

【例 2-79】 如何用 do over 语句将隐含数组的所有元素乘以 50。设程序名为 SASTJFX2_82.sas。

```

data SASTJFX2_82;
input sc01 - sc05;
array s sc01 - sc05;
do over s;          /* 等价于 do _i_ = 1 to 5 * /
s=s*50;
end;
cards;
11 14 29 12 23
;
run;

```

比较：

- (1) 简单 do 语句常作为 if-then/else 结构语句的一部分，指定执行一组语句。
- (2) 循环 do 语句基于索引变量的值，反复执行 do 和 end 语句中间的语句。
- (3) do while 语句当条件为真时，反复执行 do 语句，在每次重复执行前检查条件的真假；do until 语句在循环的末点得到条件的返回值，而 do while 语句则在循环的起点得到条件的返回值，也就是说，即使条件为假，do until 语句至少执行一次，而 do while 语句不执行。

2. end 语句

该语句停止 do 组或 select 组处理。

【例 2-80】 如何用 end 结束循环 do 组的循环处理。

```

do name = 'WANG', 'ZHANG', 'LI';
output;
end;

```

3. leave 语句

该语句停止处理当前循环，继续按顺序执行下一个语句。

【例 2-81】 如何用 leave 语句停止处理当前 do 循环。设程序名为 SASTJFX2_84.sas。

```

data SASTJFX2_84;
input name $ sex $ age weight height;
bonus=0;
do year=age TO 20;
if bonus ge 400 then leave;
bonus+100;
end;
datalines;
WANG M 14 42 150
ZHANG M 16 46 170
LI F 15 44 149

```

```
TANG M 15 38 162
LIU F 14 47 162
DIAO F 16 52 168
ZHU M 14 45 158
JIA F 16 50 167
;
run;
```

本例通过 if-then 语句检查变量 bonus 的值是否满足条件，当 bonus 的值大于等于 400 时，用 leave 语句停止处理当前 do 循环。

比较：

leave 语句停止当前循环，而 continue 语句是停止处理当前循环并继续下次循环；leave 语句可用于 do 循环和 SELECT 组，而 continue 语句只用于 do 循环结构中。

4. go to 语句

该语句指示程序立即转向标号语句规定的位置执行，如果其后有 return 语句，则返回数据步的开始位置执行。

【例 2-82】 如何用 go to 语句指示程序转向标号语句。设程序名为 SASTJFX2_85. sas。

```
data SASTJFX2_85;
  input x;
  if 1 <= x <= 5 then go to add;
  put x =;
  add: sumx + x;
datalines;
7
4
323
;
run;
```

如果 x 满足条件，则程序跳过中间其他语句，转至标号语句 add 指示的位置执行，如果 x 不满足条件，则顺序执行语句，所以当遇到观测 4 时，不执行 put 语句。

5. link 语句

该语句指定一个标号语句作为 link 语句的目的地。

【例 2-83】 如何用 link 语句执行一组语句。设程序名为 SASTJFX2_86. sas。

```
data SASTJFX2_86;
  input name $ score lesson $;
  if name = 'WANG' then link calcu;
  return;
  calcu: if lesson = 'lesson_1'
  then finalscore = score + 10;
  else if lesson = 'lesson_2'
  then finalscore = score - 5;
  return;
datalines;
WANG 56 lesson_1
ZHANG 98 lesson_2
WANG 76 lesson_2
;
run;
```

当 `name = 'WANG'` 时，程序跳转到标签语句 `calcu` 指向的目标执行，直到遇到 `return` 语句，`return` 语句将 SAS 跳转到 `link` 语句后的第 1 个语句执行，然后返回数据步开始位置读取下一个观测，当 `name` 的值不为 `WANG` 时，SAS 执行赋值语句，写入观测，并返回数据步开始位置。

新生成的数据集为：

| name | score | lesson | finalscore |
|-------|-------|----------|------------|
| WANG | 56 | lesson_1 | 66 |
| ZHANG | 98 | lesson_2 | . |
| WANG | 76 | lesson_2 | 71 |

比较：

`link` 和 `go to` 语句的区别在于其后 `return` 语句的动作，`link` 语句后的 `return` 语句返回执行 `link` 语句后的语句。`go to` 语句后的 `return` 语句则返回数据步开始位置执行，除非 `link` 语句在 `go to` 语句之前，在这种情况下，继续执行 `link` 语句后的第 1 个语句。`link` 语句中常有一个明确的 `return` 语句，而 `go to` 语句的中常无明确的 `return` 语句。

当在一个程序中需要在几个点都执行一组语句时，使用 `link` 语句能使代码更简洁、更具逻辑性；如果一组语句在一个程序中只执行一次时，使用 `do` 组语句比使用 `link - return` 语句更具有逻辑性。

6. 标号语句

该语句确定一个连接到其他语句的语句。

【例 2-84】 如何使用标号语句确定一个连接到其他语句的语句。设程序名为 `SASTJFX2_87.sas`。

```
data SASTJFX2_87A SASTJFX2_87B;
input Item $ Stock @ ;
if Stock=0 then go to reorder;
output SASTJFX2_87A;
return;
reorder: input Supplier $ ;
output SASTJFX2_87B;
datalines;
milk0 A
bread 3 B
;
run;
```

当 `Stock = 0` 时，则使用 `go to` 语句跳转到 `reorder` 标号语句指示的语句执行，即执行“`input Supplier $;`”语句和其后 `output` 语句，创建成 `SASTJFX2_87B`，否则返回数据步的开始位置继续处理下一个观测，创建成 `SASTJFX2_87A`。

7. if-then/else 语句

表达式为真时执行 `THEN` 后面的语句，表达式为假时执行 `ELSE` 后面的语句。

【例 2-85】 如果遇到 `sex = 'F'` 的观测则不写入数据集，否则将观测写入数据集。设程序名为 `SASTJFX2_88.sas`。

```

data SASTJFX2_88;
  input name $ sex $ @@ ;
  if sex='F' then delete;
  else output;
datalines;
WANG M
ZHANG M
LI F
;
run;

```

8. select 语句

该语句执行一个或几个语句或组群语句。

【例 2-86】 无选择表达式的 select 语句。设程序名为 SASTJFX2_89. sas。

```

data SASTJFX2_89;
input x;
select;
  when(x=0) x+1;
  when(x>0);
  otherwise put 'x<0';
end;
datalines;
-1
10
0
.
;
run;

```

当 $x=0$ 为真时 x 的值加 1, 当 $x>0$ 为真时不执行任何动作, 其他情况下(本例中遇到 -1 和缺失值两个观测值时)执行 put 语句。

9. return 语句

该语句停止执行数据步当前点, 返回数据步预先指定的点。

【例 2-87】 如何用 return 语句让 SAS 系统返回到数据步开头。设程序名为 SASTJFX2_90. sas。

```

data SASTJFX2_90;
  input x y;
  if x=y then return;
  put x=y=;
datalines;
21 25
20 20
7 17
;
run;

```

当遇到 $x=y$ 的观测时, SAS 执行 return 语句, 未执行 if 语句后的 put 语句, 并将观测写入数据集; 当遇到的观测不符合条件时, 执行剩余的语句, 并将观测写入数据集。

2.2 SAS 过程步中常用 SAS 语句

过程步(PROC 步)要求 SAS 从过程库中调出一个过程并执行这个过程,通常用 SAS 数据集作为输入。PROC 步以 PROC 语句开始,在 PROC 步里的其他语句给出用户想得到有关结果的更多信息的程序语句。可以用在 PROC 步的语句依赖于用户调用的这个特殊的过程。

1. var 语句

var 语句在很多过程步中用来指定被分析的变量。

【例 2-88】 如何用 var 语句指定被分析的变量。设程序名为 SASTJFX2_91.sas。

```
var xuetangzhi nianling abc tttwww;  
var x1 - x11 a1 - a99;
```

2. by 语句

by 语句在过程步中一般用来指定一个或几个分组变量,根据这些分组变量值把观测(即个体)分组,然后对每一组观测分别进行本过程指定的分析。在使用带有 by 语句的过程步之前一般要求先用 SORT 过程对数据集排序。

【例 2-89】 如何根据 by 语句指定的变量进行排序。设程序名为 SASTJFX2_92.sas。

```
proc sort data = SASTJFX2_92;  
by age;  
run;
```

3. class 语句

在某些过程步中(如 anova 或 glm),使用 class 语句指定一个或几个分类变量,这些分类变量就将充当影响因素;而在另一些过程步中(如 means),class 语句的作用与 by 语句类似,可以指定分类变量,把观测按分类变量分组后分别进行分析,此时使用 class 语句就不需要先按分类变量排序了。

【例 2-90】 如何用 class 语句指定分组变量。设程序名为 SASTJFX2_93.sas。

```
proc ttest cochrans;  
class sex;  
var weight;  
run;
```

4. model 语句

model 语句在一些统计建模过程中用来指定模型的形式,相当于以简略的形式呈现的计算公式。

【例 2-91】 如何用 model 语句指定模型的形式。

```
model xuetangzhi = nianling;
```

此语句的作用取决于它被引用的 SAS 过程,若 SAS 过程为“reg”,上面的 model 语句的含义是建立用定量的自变量 nianling 去预测定量的因变量 xuetangzhi 取值的直线回归方程,常称为直线回归分析;若 SAS 过程为“anova”,再配上“class nianling;”语句,则上面的 model 语句的含义是检验定性因素 nianling 全部水平下定量结果变量 xuetangzhi 的平均值之间的差别是否具有统计学意义,常称为单因素多水平设计定量资料方差分析。

5. freq 语句

freq 语句指定一个重复数变量，每个观测中此变量的值说明这个观测实际代表多少个完全相同的重复观测。

freq 变量只取整数值。如果变量不是整数，SAS 会将它截为整数。如果变量小于 1 或缺失，程序将不使用这些观测来计算统计量。如果没有 freq 语句，那么每个观测的默认频数为 1。频数变量的总和代表观测的总数。

【例 2-92】 如何用 freq 语句指定重复数变量。设程序名为 SASTJFX2_95. sas。

```
proc univariate noprint;
  histogram x/midpoints =90 to 160 by 10;
  freq fre;
run;
```

6. weight 语句

weight 语句指定一个权重变量，在某些允许加权的过程中代表权重。

weight 语句中变量不一定是整数，当遇到非正权重变量值时，程序的行为见表 2-4。

表 2-4 权重变量值及程序的行为

| 权重变量值 | 程 序 |
|-------|----------------------|
| 0 | 计算总观测数中的观测 |
| 小于 0 | 将权重值转为 0 并计算总观测数中的观测 |
| 缺失 | 从分析中将观测排除 |

注意：在 FREQ 过程中，weight 语句中变量的值代表每一个观测出现的频数。

【例 2-93】 如何用 weight 语句指定权重变量。设程序名为 SASTJFX2_96. sas。

```
proc catmod;
  weight f;
  model a* b* c=y;
run;
```

7. id 语句

有些过程(如 print、univariate)需要输出观测的代号，一般使用观测的序号。但是，如果数据集中有一个变量可以用来区分观测(如人名)，就可以用 id 语句指定这个变量作为观测标志。

【例 2-94】 如何用 id 语句指定作为观测标志的变量。

```
id name;
```

指定用变量 name 的值来标志观测。

8. where 语句

用 where 语句可以选择输入数据集的一个行子集来进行分析，在 where 关键字后指定一个条件。

【例 2-95】 如何用 where 语句选择子集。

```
where age >=50 and age < =70;
```

指定只分析年龄(设其变量名为 age)高于 50 且小于等于 70 岁的被调查者。

9. label 语句

label 语句为变量指定一个标签，很多过程可以使用这样的标签。

【例 2-96】 如何用 label 语句指定标签。

```
label region = 'Sales Region';
```

指定变量 region 的标签为 Sales Region。

10. output 语句

SAS 过程可以对数据集作某些分析,这时结果出现在 output 窗口(高精度绘图过程的输出在 graphics 窗口)。在 SAS 过程步中经常用 output 语句指定输出结果存放的数据集。不同过程中把输出结果存入数据集的方法各有不同,output 语句是用得最多的一种。

【例 2-97】 如何用 output 语句指定存放输出结果的数据集。

```
output out = b;
```

存放输出结果的数据集为临时数据集 b。

2.3 可在 SAS 程序中任何地方出现的 SAS 语句——全程语句

2.3.1 全程数据存取语句

1. libname 语句

该语句将 SAS 逻辑库与逻辑库引用名关联在一起或取消关联;清除一个或全部逻辑库引用名;列出 SAS 逻辑库特征;合并 SAS 逻辑库;隐含关联 SAS 目录。

【例 2-98】 如何用 libname 语句指定和使用逻辑库引用名。设程序名为 SASTJFX2_101.sas。

```
libname stulib 'd:\SASTJFX\data\';  
option user = stulib;  
data SASTJFX2_101;  
input name $ height;  
datalines;  
WANG 150  
ZHANG 170  
;  
run;
```

程序中首先定义了一个名为 stulib 的逻辑库引用名,然后通过 option 语句将当前用户的工作逻辑库命名为 stulib,然后在 stulib 逻辑库下建立名为 BCJQ2_31 的数据集。指定储存数据集物理地址时,也可以不使用 option 语句,而直接使用数据集二级名字 stulib. SASTJFX2_101。

2. filename 语句

该语句将 SAS 文件引用名(FILEREF, File Reference 的自定义缩写)同外部文件或输出驱动器连接在一起;取消文件和文件引用名的连接;列出外部文件的属性。

【例 2-99】 如何用 filename 语句为外部文件指定文件引用名。设程序名为 SASTJFX2_102.sas。

```
filename myfile 'd:\SASTJFX\2_1.txt';  
data SASTJFX2_102;  
infile myfile;  
input name $ sex $ age weight height;  
run;  
filename myfile list;
```


上面的程序中，生成 SASTJFX2_102 数据集，建立 5 个变量，从操作系统指定目录下的 2_1.txt 文件中读入观测数据。首先用 FILENAME 语句将 myfile 定义为 d:\SASTJFX\2_1.txt 文件的文件引用名，在 INFILE 语句中就可以直接使用文件引用名来引用外部文件，最后用 FILENAME 语句在日志窗口显示文件引用名 MYFILE 的物理地址。

比较：

filename 语句指定外部文件的文件引用名；libname 语句指定 SAS 数据集或可被存取 DBMS 文件的逻辑库引用名。filename 语句用于目录存取方法、剪贴板存取方法、电子邮件数据存取方法、FTP 数据存取方法等各类用法，在此不做详细介绍。

2.3.2 全程日志控制语句

1. page 语句

该语句在 SAS 日志中从新的一页开始。

【例 2-100】 如何用 page 语句将 SAS 日志从新的一页开始。

```
page;
```

用户可以在窗口环境、批处理或非交互行模式运行 SAS 过程中使用 page 语句，page 语句本身不被显示在日志中；当 SAS 以交互行模式运行时，page 语句可能显示为空格。

2. skip 语句

该语句在 SAS 日志中产生空白行。

【例 2-101】 如何用 skip 语句在 SAS 日志中产生 10 行空白。

```
skip 10;
```

2.3.3 全程环境控制语句

X 语句

该语句在 SAS 会话中发布操作系统命令。

【例 2-102】 如何用 X 语句向操作系统发布命令。设程序名为 SASTJFX2_105.sas。

```
data SASTJFX2_105;
infile 'd:\SASTJFX\2_1.txt';
input name $ sex $ age weight height;
run;
X 'del d:\SASTJFX\2_1.txt';
```

单引号中为 DOS 命令，意为删除计算机 d 盘指定目录下的 2_1.txt 文件。

2.3.4 全局输出控制语句

1. footnote 语句

该语句在过程步或程序步输出页的底部写入不超过 10 行的文本。

【例 2-103】 如何用 footnote 语句规定脚注。

```
footnote8 'Managers' Meeting';
```

用 footnote 语句规定在脚注的第 8 行写入一个文本信息。

2. title 语句

该语句给 SAS 输出指定一个标题行。

【例 2-104】 如何用 title 语句规定标题。

```
title5 color = red ''Year's End Report'';
```

设置输出行的标题''Year's End Report'', 且标题位于第 5 行, 字体为红色。

2.3.5 全程序控制语句

1. dm 语句

该语句像一条 SAS 语句一样, 提交 SAS 程序编辑器、日志、过程输出或文本编辑器命令。

【例 2-105】 如何用 dm 语句改变窗口颜色。

```
dm 'color text cyan;color command red';
```

dm 语句将程序编辑器窗口的文本颜色改为深蓝色, 命令行颜色为红色。

【例 2-106】 如何用 dm 语句清除日志窗口内容, 并使结果输出窗口成为活动窗口:

```
dm log 'clear' output;
```

2. endsas 语句

当执行完数据步或过程步后, 该语句结束 SAS 作业或对话。

endsas 语句常用于交互式或窗体操作会话中, 它不能用于诸如 if-then 的选择结构中。

【例 2-107】 如何用 endsas 语句退出 SAS 系统。

```
endsas;  
run;
```

3. %include 语句

该语句将 SAS 程序语句和(或)数据行带入当前 SAS 程序。

【例 2-108】 如何用 %include 语句调用外部文件。

```
%include 'd:\SASTJFX\2_1.txt';
```

比较:

%include 语句立即执行, 而 include 命令带入输入行到程序编辑器窗口并不执行, 需要有一个提交运行命令来执行。

4. lock 语句

该语句获得和释放对已存在的 SAS 文件的独占锁定。

【例 2-109】 如何用 lock 语句对 work.SASTJFX2_112 数据集获得独占访问锁定。

```
lock work.SASTJFX2_112;
```

5. options 语句

该语句改变一个或多个 SAS 系统选项的值。

【例 2-110】 如何用 options 语句设置 SAS 系统选项的值。

```
options nodate linesize = 72;
```

本例在结果输出中不显示日期，输出结果页面为 72 行。

比较：

改变 SAS 系统选项可以通过 options 语句，或显示管理命令 options。用显示管理命令时，选项名和设置都在窗口相应的列上显示，为改变某种设置，只需简单地在显示的值上按 Enter 键。

6. run 语句

该语句执行先前输入的 SAS 语句。

7. quit 语句

该语句结束一个交互式过程。

【例 2-111】 如何用 quit 语句结束交互式过程。设程序名为 SASTJFX2_114.sas。

```
proc gplot data=a;  
plot chpr* data;  
title 'First plot';  
run;  
plot dekl* data;  
title 'Second plot';  
run;  
quit;
```

比较：

quit 语句用来结束 gplot 过程，因为 gplot 过程为交互式的，run 语句不能结束该过程，它只是告诉 gplot 去执行 run 语句之前的语句。

2.4 SAS 函数中的基础知识

SAS 像多数高级语言一样，提供了丰富的标准函数，并且在内容上不断地发展改进，数量也从 SAS 6.12 的 200 多个增加到 SAS 9.2 的近 500 个。本章列举了最常用的 7 类 SAS 函数和 SAS call 子程序，以及应用实例，有助于读者学习和使用。一旦学会使用合适的函数，将会使用户的工作效率加倍。

2.4.1 SAS 函数

SAS 函数是一个子程序，它由 0 或几个参数返回一个结果值。每个 SAS 函数都有一个关键词名字。为了引用函数，要写出它的名字，然后写出一个参数或几个参数，将其放入括号中，而后这个函数对这些参数进行某种运算。

SAS 函数书写格式为：

函数名(参数表)

其中，函数名为 SAS 关键字，参数表给出函数所要求的一个或多个参数。

2.4.2 SAS 参数

SAS 参数可以是变量名、常数、函数或表达式。

当参数多于一个时，参数之间应该用逗号分隔。例如可以用函数 SUM(x1,x2,x3) 求 x1, x2, x3 三个自变量的和。

也可以写成“函数名(OF 变量名列表)”格式。其中，变量名列表可以是任何合法的变量名列表。比如 x1, x2, x3 的和等价地可以用 SUM(of x1 x2 x3)或 SUM(of x1 - x3)表示。

但需要注意的是两种写法不能混在一起，比如 SUM(of x1 ,x2 ,x3)和 SUM(x1 - x3)都是不正确的。

2.4.3 函数值

除了个别特殊情况外，多数函数值的属性与其参数属性是一致的，数值函数值的默认长度为 8 个字节，字符函数值的默认长度是 200 个字符。

2.4.4 SAS 函数分类

最新 SAS 9.2/BASE 模块的 Language Reference: Dictionary 中的内置函数，与 SAS 9.1.3 比较，减少了程序响应度量函数(ARM)、货币折算函数(Currency Conversion)、双字节函数(DBCS)等函数种类，新增了算术函数(Arithmetic)、组合函数(Combinatorial)、距离函数(Distance)、财务函数(Finance)、数字函数(Numeric)、搜索函数(Search)、排序函数(Sort)等，具体分类见表 2-5。

表 2-5 SAS 函数分类

| 分 类 | 中文名称 | 分 类 | 中文名称 |
|----------------------------|---------|----------------------|---------------|
| Arithmetic | 算术函数 | Numeric | 数字函数 |
| Array | 数组函数 | Probability | 概率函数 |
| Bitwise Logical Operations | 逐位逻辑函数 | Quantile | 分位数函数 |
| Character String Matching | 字符串匹配函数 | Random Number | 随机数函数 |
| Character | 字符函数 | SAS File I/O | SAS 文件输入/输出函数 |
| Combinatorial | 组合函数 | Search | 搜索函数 |
| Date and Time | 日期时间函数 | Sort | 排序函数 |
| Descriptive Statistics | 样本统计函数 | Special | 特殊函数 |
| Distance | 距离函数 | State and Zip Code | 州和 ZIP 码换算函数 |
| External Files | 外部文件函数 | Trigonometric | 三角函数 |
| External Routines | 外部程序函数 | Truncation | 截取函数 |
| Financial | 财务函数 | Variable Control | 变量控制函数 |
| Hyperbolic | 双曲线函数 | Variable Information | 变量信息函数 |
| Macro | 宏函数 | Web Tools | 网络工具 |
| Mathematical | 数学函数 | | |

2.4.5 SAS 函数在使用中的注意事项

- (1) SAS 函数不能直接用在 PUT 语句中。
- (2) 多数函数不许以缺失值为其参数。
- (3) 注意函数参数的取值范围，例如对数函数的参数值要大于零等。
- (4) 若某些概率函数的参数选择不当，则可能引起收敛性问题。在这种情况下，函数值为缺失值并给出错误信息。

2.5 日期时间函数

2.5.1 日期时间函数简介

日期时间函数的内容很多，见表 2-6。

表 2-6 日期时间函数

| 函 数 | 描 述 |
|---|---|
| DATDIF(sdate, edate, basis) | 计算两个日期之间相距的天数，basis 指定 SAS 中的时间格式 |
| DATE() | 计算当前日期作为 SAS 日期数据 |
| DATEJUL(julian-date) | 将 julian 西洋日期格式转换为 SAS 日期格式 |
| DATEPART(datetime) | 从日期时间格式的数据中抽取日期 |
| DATETIME() | 计算当前日期时间值 |
| DAY(date) | 由 SAS 日期值 date 得到日 |
| DHMS(date, hour, minute, second) | 从日期小时分钟秒四个数值得到 SAS 日期时间值 |
| HMS(hour, minute, second) | 由小时分钟秒三个值计算 SAS 日期时间值 |
| HOLIDAY(holiday, year) | 计算指定年份指定节日的 SAS 日期时间值 |
| HOUR(time datetime) | 从 SAS 时间或 SAS 日期时间中计算小时的数值 |
| INTCINDEX(interval << multiple. < shift-index >>>, date-time-value) | 按给定日期、时间或日期时间值，计算周期指数 |
| INTCK('interval', from, to) | 计算在一定时间间隔后的 SAS 日期时间值 |
| INTCYCLE(interval << multiple. < shift-index >>>) | 按给定日期、时间或日期时间间隔，返回下一较高季节周期的日期、时间或日期时间间隔 |
| INTFIT(argument-1, argument-2, 'type') | 返回两个日期之间的时间间隔 |
| INTFMT(interval << multiple. < . shift-index >>>, 'size') | 按给定日期、时间或日期时间间隔，返回推荐的格式 |
| INTGET(date-1, date-2, date-3) | 返回基于三个日期值或日期时间值的时间间隔 |
| INTINDEX(interval << multiple. < shift-index >>>, date-value) | 按给定日期、时间或日期间隔，计算周期指数 |
| INTNX(interval, from, n) | 计算从 from 开始经过 n 个间隔后的 SAS 日期时间值 |
| INTSEAS(interval << multiple. < shift-index >>>) | 按给定日期、时间或日期时间间隔，返回季节周期的长度 |
| INTSHIFT(interval << multiple. < shift-index >>>) | 返回与基时间间隔相对应的移位时间间隔 |
| INTTEST(interval << multiple. < shift-index >>>) | 若时间间隔有效，返回 1；若时间间隔无效，则返回 0 |
| JULDATE(date) | 将 SAS 日期格式转换为 julian 西洋日期 |
| MDY(month, day, year) | 从月日年得到一个 SAS 日期值 |
| MINUTE(time datetime) | 从时间值或日期时间值中抽取分钟值 |
| MONTH(date) | 从日期中得到月份 |
| NWKDOM(n, weekday, month, year) | 返回指定年的指定月中某个星期几第 n 次出现时的日期 |
| QTR(date) | 从 SAS 日期值得到对应的季度 |
| SECOND(time datetime) | 从 SAS 时间值或 SAS 日期时间值中得到秒 |
| TIME() | 计算当前时间 |
| TIMEPART(datetime) | 从日期时间值中抽取时间值 |
| TODAY() | 计算当前日期 |
| WEEK(date, < descriptor >) | 计算周数 |
| WEEKDAY(date) | 由 SAS 日期值得到一周内的第几天 |
| YEAR(date) | 由 SAS 日期值得到年份 |
| YRDIF(sdate, edate, basis) | 计算两个日期之间所差的年份 |
| YYQ(year, quarter) | 从年份和季度产生一个 SAS 日期值 |

2.5.2 用 DATDIF 函数计算两个日期之间的天数

【例 2-112】 求从 2011 年 1 月 1 日至 2012 年 1 月 1 日之间的天数(按每个月 30 天计算)。
具体 SAS 程序如下, 设程序名为 SASTJFX2_115. sas。

| 程序 | 说明 |
|---|-----------------------------|
| data SASTJFX2_115; sdate = '01jan2011'd; edate = '01jan2012'd; days = datdif(sdate,edate, '30/360'); put days = ; run; | 调用 datdif 函数 按每个月 30 天计算 |

结果显示为:

```
days = 360
```

【例 2-113】 求从 1982 年 7 月 30 日至 2012 年 5 月 6 日之间的天数(按每个月实际天数计算)。

具体 SAS 程序如下, 设程序名为 SASTJFX2_116. sas。

| 程序 | 说明 |
|---|----------------------------|
| data SASTJFX2_116; sdate = '30jul1982'd; edate = '06may2012'd; days = datdif(sdate,edate, 'act/act'); put days = ; run; | 调用 datdif 函数 按每个月实际天数计算 |

结果显示为:

```
days = 10873
```

2.5.3 用 YRDIF 函数计算两个日期之间的年数

【例 2-114】 求从 2001 年 1 月 1 日至 2011 年 1 月 1 日之间的年数。

具体 SAS 程序如下, 设程序名为 SASTJFX2_117. sas。

| 程序 | 说明 |
|--|--|
| data SASTJFX2_117; sdate = '01jan2001'd; edate = '01jan2011'd; year1 = yrdif(sdate,edate, '30/360'); year2 = yrdif(sdate,edate, 'act/360'); put year1 = year2 = ; run; | 调用 yrdif 函数 按一年 360 天、每个月 30 天计算 按一年 360 天、每个月实际天数计算 |

结果显示为:

```
year1 = 10  year2 = 10.144444444
```

【例 2-115】 求从 2008 年 8 月 26 日至 2013 年 6 月 10 日之间的年数。

具体 SAS 程序如下, 设程序名为 SASTJFX2_118. sas。

| 程序 | 说明 |
|--|---|
| <pre>data SASTJFX2_118; sdate = '26aug2008'd; edate = '10jun2013'd; year1 = yrdif(sdate, edate, 'act/act'); year2 = yrdif(sdate, edate, 'act/365'); put year1 = year2 = ; run;</pre> | 调用 yrdif 函数 按实际天数计算 按一年 365 天、每个月实际天数计算 |

结果显示为:

```
year1 = 4.7880829403   year2 = 4.7917808219
```

2.5.4 用 HOUR 和 MINUTE 函数计算当前时间

【例 2-116】 计算当前时间。

具体 SAS 程序如下, 设程序名为 SASTJFX2_119. sas。

| 程序 | 说明 |
|--|-----------------------|
| <pre>data SASTJFX2_119; x = hour(datetime()); y = minute(datetime()); put x = y = ; run;</pre> | 调用 hour 和 minute 两个函数 |

结果显示为:

```
x = 17   y = 16, 运行该段程序时间为当天第 17 个小时第 16 分钟。
```

2.5.5 用 YEAR、QTR、MONTH 和 DAY 函数计算当前所处的年、季度、月份和日期

【例 2-117】 计算当前所处的年、季度、月份和日期。

具体 SAS 程序如下, 设程序名为 SASTJFX2_120. sas。

| 程序 | 说明 |
|---|------------------------------|
| <pre>data SASTJFX2_120; year = year(date()); qtr = qtr(date()); month = month(date()); day = day(date()); put year = qtr = month = day = ; run;</pre> | 调用 year、qtr、month 和 day 四个函数 |

结果显示为:

```
year = 2011   qtr = 4   month = 12   day = 8, 运行该段程序时间为 2011 年第 4 季度 12 月 8 日。
```

2.5.6 用 HOLIDAY 函数计算指定年份指定节日的日期

【例 2-118】 计算 2012 年圣诞节的日期。

具体 SAS 程序如下，设程序名为 SASTJFX2_121. sas。

| 程序 | 说明 |
|--|---|
| <pre>data SASTJFX2_121; holiday=holiday('Christmas',2012); put holiday=WEEKDATE.; run;</pre> | 调用 holiday 函数 输出格式为 WEEKDATE. (输出星期和日期值) |

结果显示为：

holiday = Tuesday, December 25, 2012

2.6 截取函数

2.6.1 截取函数简介

截取函数见表 2-7。

表 2-7 截取函数

| 函数 | 描述 | 函数 | 描述 |
|----------|--------------------------------|------------|-----------------|
| CEIL(x) | 取大于等于变量 x 的最小整数 | INT(x) | 取 x 的整数部分 |
| FLOOR(x) | 取小于等于变量 x 的最大整数 | ROUND(x,n) | x 按 n 指定的精度取舍入值 |
| FUZZ(x) | 当自变量 x 和某个整数的差在 1E-12 的范围内时取整数 | TRUNC(x,q) | x 按规定长度 q 截取的数值 |

2.6.2 用 CEIL 函数求最小整数

【例 2-119】 求大于等于 25.13 的最小整数。

具体 SAS 程序如下，设程序名为 SASTJFX2_122. sas。

| 程序 | 说明 |
|---|------------|
| <pre>data SASTJFX2_122; x=ceil(25.13); put x=; run;</pre> | 调用 ceil 函数 |

结果显示为：

x = 26

2.6.3 用 FLOOR 函数求最大整数

【例 2-120】 求不超过 -135.2 的最大整数。

具体 SAS 程序如下，设程序名为 SASTJFX2_123. sas。

| 程序 | 说明 |
|---|--------------|
| <pre>data SASTJFX2_123; x=floor(-135.2); put x=; run;</pre> | 调用 floorl 函数 |

结果显示为：

x = -136

2.6.4 用 INT 函数取整数部分

【例 2-121】 求 -135.2 的整数部分。

具体 SAS 程序如下，设程序名为 SASTJFX2_124.sas。

| 程序 | 说明 |
|---|-----------|
| <pre>data SASTJFX2_124; x=int(-135.2); put x=; run;</pre> | 调用 int 函数 |

结果显示为：

$x = -135$

2.6.5 用 ROUND 函数按指定的精度取舍入值

【例 2-122】 将 2012.372 精确到百分位数。

具体 SAS 程序如下，设程序名为 SASTJFX2_125.sas。

| 程序 | 说明 |
|--|-------------|
| <pre>data SASTJFX2_125; x=round(2012.372,0.01); put x=; run;</pre> | 调用 round 函数 |

结果显示为：

$x = 2012.37$

2.6.6 用 TRUNC 函数求截取数值

【例 2-123】 求 $1/5$ 按 3 个字节存储时的值。

具体 SAS 程序如下，设程序名为 SASTJFX2_126.sas。

| 程序 | 说明 |
|--|-------------|
| <pre>data SASTJFX2_126; x=trunc(1/5,3); put x=; run;</pre> | 调用 trunc 函数 |

结果显示为：

$x = 0.1999816895$

2.7 分位数函数

2.7.1 分位数函数简介

设连续型变量 X 的分布函数为 $F(x)$ ，对给定的 $p(0 \leq p \leq 1)$ ，若有 x_p 使得 $F(x_p) = p$ ，则称 x_p 为随机变量 X 的 p 分位数(或称分布 $F(x)$ 的 p 分位数)。SAS 中常见分位数列表见表 2.8。

表 2-8 分位数函数

| 函 数 | 描 述 | 函 数 | 描 述 |
|--------------------|-------------|--|--------------|
| BETAINV(p,a,b) | 计算贝塔分布的分位数 | PROBIT(p) | 计算标准正态分布的分位数 |
| CINV(p,df,nc) | 计算卡方分布的分位数 | QUANTILE('dist',probability, parm-1,...,parm-k) | 计算指定分布的分位数 |
| FINV(p,ndf,ddf,nc) | 计算 F 分布的分位数 | TINV(p,df,nc) | 计算 t 分布的分位数 |
| GAMINV(p,a) | 计算伽马分布的分位数 | | |

2.7.2 用 CINV 函数计算卡方分布曲线下的 p 分位数

χ^2 分布分位数

`cinv(p,df,nc)` ($0 \leq p \leq 1, df > 0, nc \geq 0$)

该函数计算自由度为 `df`，非中心参数为 `nc` 的 χ^2 分布的 `p` 分位数。取 `nc = 0` 或不规定此项参数表明是中心 χ^2 分布。

【例 2-124】 计算自由度为 3.5，非中心参数为 4.5 的 χ^2 分布的 0.95 分位数。
具体 SAS 程序如下，设程序名为 `SASTJFX2_127.sas`。

| 程序 | 说明 |
|--|-------------------------|
| <pre>data SASTJFX2_127; q=cinv(0.95,3.5,4.5); put q=; run;</pre> | 调用 <code>cinv</code> 函数 |

结果显示为：
`q = 17.504582117`

2.7.3 用 FINV 函数计算 F 分布曲线下的 p 分位数

F 分布的分位数

`finv(p,ndf,ddf,nc)` ($0 \leq p \leq 1, ndf > 0, ddf > 0, nc \geq 0$)

该函数计算分子自由度为 `ndf`，分母自由度为 `ddf`，非中心参数为 `nc` 的 F 分布的 `p` 分位数。若取 `nc = 0` 或没有规定 `nc`，它就是中心 F 分布的 `p` 分位数。若 `nc` 太大，那么使用的算法可能不成功，这种情况函数得到一个缺失值。

【例 2-125】 计算分子自由度为 2，分母自由度为 10，没有规定非中心参数的 F 分布的 0.95 分位数。
具体 SAS 程序如下，设程序名为 `SASTJFX2_128.sas`。

| 程序 | 说明 |
|---|-------------------------|
| <pre>data SASTJFX2_128; q=finv(0.95,2,10); put q=; run;</pre> | 调用 <code>finv</code> 函数 |

结果显示为：

```
q = 4.1028210151
```

2.7.4 用 PROBIT 函数计算标准正态分布曲线下的 p 分位数

正态分布的分位数

$\text{probit}(p) (0 \leq p \leq 1)$

该函数计算标准正态分布函数的分位数。它是概率函数 probnorm 的逆函数。如果随机变量 $X \sim N(0,1)$ ，则

$P\{X \leq \text{probit}(z)\} = z$

这个函数产生的结果在 -5 和 5 之间。

【例 2-126】 计算标准正态分布函数的 0.95 分位数。

具体 SAS 程序如下，设程序名为 SASTJFX2_129.sas。

| 程序 | 说明 |
|--|--------------|
| data SASTJFX2_129; q=probit(0.95); put q=; run; | 调用 probit 函数 |

结果显示为：

```
q = 1.644853627
```

2.7.5 用 TINV 函数计算 t 分布曲线下的 p 分位数

t 分布的分位数

$\text{tinv}(p, df, nc) (0 \leq p \leq 1, df > 0)$

该函数计算自由度为 df ，非中心参数为 nc 的 t 分布的 p 分位数。若 nc 没有规定或取 $nc = 0$ ，那么计算的就是中心 t 分布的分位数。若 nc 的绝对值很大，使用的算法可能失败。这种情况函数得到一个缺失值。

【例 2-127】 计算自由度为 3，非中心参数为 5 的 t 分布的 0.95 分位数。

具体 SAS 程序如下，设程序名为 SASTJFX2_130.sas。

| 程序 | 说明 |
|--|------------|
| data SASTJFX2_130; q=tinv(0.95,3,5); put q=; run; | 调用 tinv 函数 |

结果显示为：

```
q = 15.066410178
```

2.8 数学函数

2.8.1 数学函数简介

数学函数见表 2-9。

表 2-9 数学函数

| 函 数 | 描 述 |
|--|---------------------|
| ABS(x) | 求绝对值 |
| AIRY(x) | 计算 AIRY 函数值 |
| BETA(a,b) | 计算 BETA 函数值 |
| CNONCT(x,df, prob) | 求卡方分布的非中心参数值 |
| COALESCE(x1,x2,...,xn) | 求下列数值的第一个非缺失值 |
| DAIRY(x) | 计算 AIRX 函数的导数 |
| DEVIANC(distribution,variable,shape-parameter(s), ε) | 计算基于概率分布的偏差 |
| DIGAMMA(x) | 计算伽马函数的对数值 |
| ERF(x) | 计算误差函数 |
| ERFC(x) | 计算误差函数的余函数 |
| EXP(x) | 计算指数函数值 |
| FACT(n) | 计算阶乘 |
| FNONCT(x,ndf,ddf,prob) | 求 F 分布的非中心参数值 |
| GAMMA(x) | 计算完全伽马函数值 |
| GCD(x1, x2, x3, ..., xn) | 计算一个或多个整数的最大公约数 |
| IBESSEL(nu,x,kode) | 计算修正的 bessell 函数 |
| JBESSEL(nu,x) | 计算 bessell 函数值 |
| LCM(x1, x2, x3, ..., xn) | 计算能被一组数中的每个数整除的最小倍数 |
| LGAMMA(x) | 计算伽马函数的自然对数 |
| LOG(x) | 计算自然对数 |
| LOG1PX(x) | 计算 1 加该参数的对数 |
| LOG10(x) | 对自变量 x 求以 10 为底的对数 |
| LOG2(x) | 对自变量 x 求以 2 为底的对数 |
| LOGBETA(a,b) | 计算 beta 函数的对数 |
| MOD(x1,x2) | 计算余数值 |
| SIGN(x) | 返回数字正负 1 或 0 |
| SQRT(x) | 计算算数平方根 |
| TNONCT(x,df,prob) | 求 t 分布的非中心参数值 |
| TRIGAMMA(x) | 计算 TRIGAMMA 函数值 |

2.8.2 用 ABS 函数求绝对值

【例 2-128】 求 -171 的绝对值。

具体 SAS 程序如下，设程序名为 SASTJFX2_131.sas。

| 程序 | 说明 |
|--|-----------|
| data SASTJFX2_131; x = abs(-171); put x =; run; | 调用 abs 函数 |

结果显示为：

```
x = 171
```

2.8.3 用 EXP 函数计算 e 的 x 次幂

【例 2-129】 计算 $e^{3.14}$ 的值。

具体 SAS 程序如下，设程序名为 SASTJFX2_132.sas。

| 程序 | 说明 |
|--|-----------|
| data SASTJFX2_132; x = exp(3.14); put x =; run; | 调用 exp 函数 |

结果显示为：

```
x = 23.103866859
```

2.8.4 用 LOG 函数计算以 e 为底的真数 x 的自然对数值

【例 2-130】 计算 $\log(5)$ 的值。

具体 SAS 程序如下，设程序名为 SASTJFX2_133.sas。

| 程序 | 说明 |
|---|-----------|
| data SASTJFX2_133; x = log(5); put x =; run; | 调用 log 函数 |

结果显示为：

```
x = 1.6094379124
```

2.8.5 用 LOG10 函数计算以 10 为底的真数 x 的对数值

【例 2-131】 计算 \log_{10}^{314} 的值。

具体 SAS 程序如下，设程序名为 SASTJFX2_134.sas。

| 程序 | 说明 |
|---|-------------|
| data SASTJFX2_134; x = log10(314); put x =; run; | 调用 log10 函数 |

结果显示为：

```
x = 2.4969296481
```

2.8.6 用 MOD 函数计算余数值

【例 2-132】 计算 52 除以 6 商的余数。

具体 SAS 程序如下，设程序名为 SASTJFX2_135.sas。

| 程序 | 说明 |
|---|-----------|
| data SASTJFX2_135;结果显示为: x = mod(52,6);结果显示为: put x = ;结果显示为: run; | 调用 mod 函数 |

结果显示为:

x = 4

2.8.7 用 SQRT 函数计算平方根

【例 2-133】 计算 9 的平方根。

具体 SAS 程序如下，设程序名为 SASTJFX2_136. sas。

| 程序 | 说明 |
|---|------------|
| data SASTJFX2_136; x = sqrt(9); put x = ; run; | 调用 sqrt 函数 |

结果显示为:

x = 3

2.8.8 用 SQRT 函数、FNONCT 函数和 FINV 函数计算 ψ 值

ψ 值是使方差分析同时满足下面两个条件时的非中心 F 分布的非中心参数值 δ 除以试验因素的自由度 v_1 (即检验统计量 F 的分子的自由度)后开算术平方根的结果, 即 $\psi = \sqrt{\delta/v_1}$ 。两个条件为:

- (1)拒绝 H_0 时可能犯错误的最大概率为 α ;
- (2)不能拒绝 H_0 时, 若接受 H_0 , 可能犯错误的最大概率为 β 。

分子和分母的自由度分别为 ndf 和 ddf、分布曲线下左侧概率为 $1-\alpha$ 的中心 F 分布对应的分位数为 FINV($1-\alpha$,ndf,ddf), 那么, 可推出

$$\psi = \sqrt{\delta/v_1} = \sqrt{\delta/ndf} = \sqrt{FNONCT(FINV(1-\alpha,ndf,ddf),ndf,ddf,\beta)/ndf}。$$

【例 2-134】 当 $\alpha = 0.05$, ndf = 2, ddf = 10, $\beta = 0.10$ 时, 计算 ψ 值。

具体 SAS 程序如下，设程序名为 SASTJFX2_137. sas。

| 程序 | 说明 |
|---|-------------------------------|
| data SASTJFX2_137; x = sqrt(fnonct(finv(1-0.05, 2,10),2,10,0.10)/2); put x = ; run; | 调用 sqrt 函数、fnonct 函数和 finv 函数 |

结果显示为:

x = 2.9528593959

2.8.9 用 CNONCT 函数和 CINV 函数计算 λ 值

λ 值是使 χ^2 检验同时满足 2.8.8 节所列出的两个条件时的非中心 χ^2 分布的非中心参数值。

自由度为 ν 、分布曲线下左侧概率为 $1 - \alpha$ 的中心 χ^2 分布的分位数为 $\text{CINV}(1 - \alpha, \nu)$ 。可推出自由度为 ν 、分位数为 $\text{CINV}(1 - \alpha, \nu)$ 、分布曲线下左侧概率为 β 的非中心 χ^2 分布的非中心参数值 $\delta = \text{CNONCT}(\text{CINV}(1 - \alpha, \nu), \nu, \beta)$ 。该非中心参数值就是所要计算的 λ 值, 即 $\lambda = \delta$ 。

【例 2-135】 当 $\alpha = 0.05$, $\beta = 0.10$, $\nu = 4$ 时, 计算 λ 值。

具体 SAS 程序如下, 设程序名为 SASTJFX2_138.sas。

| 程序 | 说明 |
|---|-----------------------|
| <pre>data SASTJFX2_138; x=cnonct(cinv(1-0.05,4),4,0.10); put x=; run;</pre> | 调用 cnonct 函数和 cinv 函数 |

结果显示为:

x = 15.405051859

2.9 概率函数

2.9.1 概率函数简介

概率函数见表 2-10。

表 2-10 概率函数

| 函 数 | 描 述 |
|---|----------------------|
| CDF('dist', quantile, parm-1, ..., parm-k) | 计算累计分布函数 |
| LOGCDF('dist', quantile, parm-1, ..., parm-k)_ | 计算左侧累积分布函数的对数值 |
| LOGPDF('dist', quantile, parm-1, ..., parm-k)_ | 计算概率密度函数的对数值 |
| LOGSDF('dist', quantile, parm-1, ..., parm-k) | 计算生存函数的对数值 |
| PDF('dist', quantile, parm-1, ..., parm-k) | 计算概率密度函数 |
| POISSON(lambda, n) | 计算泊松分布的概率 |
| PROBBETA(x, a, b) | 计算贝塔分布的概率 |
| PROBBNML(p, n, m) | 计算二项式分布的概率 |
| PROBBNRM(x, y, r) | 计算二项正态分布的概率 |
| PROBCHI(x, df, nc) | 计算卡方分布的概率 |
| PROBF(x, ndf, ddf, nc) | 计算 F 分布的概率 |
| PROBGAM(x, a) | 计算伽马分布的概率 |
| PROBHYP(n, k, n, x, or) | 计算超几何分布的概率 |
| PROBMC(distribution, q, prob, df, nparms, parameters) | 由多组均值的多重比较分布计算概率和分位数 |
| PROBNEGB(p, n, m)_ | 计算负二项分布的概率 |
| PROBNORM(x) | 计算标准正态分布的概率 |
| PROBT(x, df, nc)_ | 计算 t 分布的概率 |
| SDF('dist', quantile, parm-1, ..., parm-k) | 计算生存函数 |

2.9.2 用 PROBCHI 函数计算服从卡方分布的随机变量小于 x 的概率

χ^2 分布的分布函数

probchi(x,df,nc)

该函数计算服从自由度为 df，非中心参数为 nc 的 χ^2 分布的随机变量小于给定 x 的事件的概率。如果 nc 没有规定或取为 0，那么被计算的就是中心 χ^2 分布。自由度 df 允许不是整数。

【例 2-136】 计算自由度为 5，中心卡方分布曲线下卡方值小于 20 的概率值。

具体 SAS 程序如下，设程序名为 SASTJFX2_139.sas。

| 程序 | 说明 |
|---|---------------|
| <pre>data SASTJFX2_139; p=probchi(20,5); put p=; run;</pre> | 调用 probchi 函数 |

结果显示为：

p=0.9987502694

2.9.3 用 PROBF 函数计算服从 F 分布的随机变量小于 x 的概率

F 分布的分布函数

probf(x,ndf,ddf,nc)

该函数计算服从分子自由度为 ndf，分母自由度为 ddf 的 F 分布的随机变量小于给定值 x 的事件的概率。自变量 nc 是中心函数。当分布是中心 F 分布时，取 nc=0 或不规定这项自变量。自由度可以是非整数。

【例 2-137】 计算自由度为 20 和 5，非中心参数为 15 的 F 分布曲线下 F 值小于 5 的概率。

具体 SAS 程序如下，设程序名为 SASTJFX2_140.sas。

| 程序 | 说明 |
|--|-------------|
| <pre>data SASTJFX2_140; p=probf(5,20,5,15); put p=; run;</pre> | 调用 probf 函数 |

结果显示为：

p=0.8772112982

2.9.4 用 PROBNORM 函数计算标准正态分布曲线下的面积

标准正态分布函数

probnorm(x)

该函数计算服从标准正态分布的随机变量 U 小于给定 x 的概率, 即 $P\{U < x\}$ 。这个函数等价于

$$\frac{1}{2} \left(1 + \operatorname{ERF} \left(\frac{x}{\sqrt{2}} \right) \right)$$

其中 ERF 是误差函数。

【例 2-138】 计算三个特殊的正态概率值, 即正态曲线下从 $-\infty$ 到给定数值之间的面积。具体 SAS 程序如下, 设程序名为 SASTJFX2_141.sas。

| 程序 | 说明 |
|--|----------------|
| <pre>data SASTJFX2_141; x=probnorm(0); y=probnorm(1.96); z=probnorm(2.576); put x=y=z=; run;</pre> | 调用 probnorm 函数 |

结果显示为:

$x=0.5 \quad y=0.9750021049 \quad z=0.9950024677$

即正态曲线下从 $-\infty$ 到 0.5、1.96 和 2.576 之间的面积分别为 0.5、0.975 和 0.995。

2.9.5 用 PROBT 函数计算服从 t 分布的随机变量小于 x 的概率

t 分布的分布函数

`probt(x, df, nc)`

该函数计算服从分子自由度为 df , 非中心参数为 nc 的 t 分布随机变量小于给定值 x 的事件的概率。若参数 nc 没有规定或取为 0, 那么被计算的就是中心 t 分布。自由度 df 允许非整数, 用 t 分布作双边检验时, 用 $(1 - \operatorname{probt}(\operatorname{abs}(x), df)) * 2$ 计算显著水平。

【例 2-139】 计算自由度为 3 的中心 t 分布曲线下 t 的绝对值大于 2.76 的概率。

具体 SAS 程序如下, 设程序名为 SASTJFX2_142.sas。

| 程序 | 说明 |
|--|-------------|
| <pre>data SASTJFX2_142; p=(1-probt(abs(2.76),3))*2; put p=; run;</pre> | 调用 probt 函数 |

结果显示为:

$p=0.0701523186$

2.9.6 用 PROBMC 函数计算 q 临界值

多组均值的多重比较分布函数

`PROBMC(distribution, q, prob, df, nparms, parameters)`

该函数计算作多个均数两两比较或多重比较时所构造的特定统计量(如学生化极差)的尾端概率或分位数。

distribution 是一个标识分布类型的字符串。当 q 等于某个具体的数值时(此时, 参数 prob 的取值必须用 \cdot 代替), 其表示服从参数 distribution 所代表的某种分布的统计量的值; 当 q 的取值用 \cdot 代替时(此时, 参数 prob 必须等于一个位于 $(0, 1)$ 之间的具体的数值), 表示将通过 PROBMCMC 函数求取参数 distribution 所代表的某种分布曲线下左侧概率为 prob 的分位数。prob 是相应分布的概率密度曲线下随机变量的取值位于特定分位数左侧的概率。 q 与 prob 必须有而且只能有一个的取值用 \cdot 来代替, 而 PROBMCMC 函数所返回的值正是 q 和 prob 两个参数中取值用 \cdot 来代替的那个参数的值。df 为方差分析时误差项的自由度, nparms 是处理组的组数。parameters 是一个可选项, 该选项为一个序列, 组成了“nparms”这个参数的具体内容。当试验因素各水平组的样本含量不等时, 必须指定该选项。nparms 这个参数的含义取决于分布的类型。若不指定这些参数, 则意味着假定各组的样本含量相等。

对学生化极差, 当自由度 $df \neq \infty$ 且各组的样本含量不等时, 函数 PROBMCMC 无效。

对 Williams 检验, 当各组的样本含量不等时, 函数 PROBMCMC 也无效。

【例 2-140】 设有某项单因素 5 水平设计, 每个水平组做 4 次独立重复试验, 测量某项定量指标的取值, 并假定资料满足方差分析的前提条件, 取检验水准 $\alpha = 0.05$, 经单因素 5 水平设计一元定量资料的方差分析处理, 发现试验因素对试验结果的影响有统计学意义, 现欲采用 Student-Newman-Keuls (SNK) 检验(即 q 检验)进行 5 个总体均值之间的多重比较, 请计算所需的 q 临界值。

由题意可知, 检验水准 $\alpha = 0.05$, 试验因素的水平数 $k = 5$, 各水平组的样本含量 $n = 4$, 故多重比较时的自由度 $df = k(n - 1) = 15$ 。所需求取的 q 临界值为 $q(5, 15, 0.05)$ 、 $q(4, 15, 0.05)$ 、 $q(3, 15, 0.05)$ 和 $q(2, 15, 0.05)$, 共 4 个。可用下面的 SAS 程序计算出上述 q 临界值。

具体 SAS 程序如下, 设程序名为 SASTJFX2_143.sas。

| 程序 | 说明 |
|---|---------------|
| <pre>data SASTJFX2_143; alpha=0.05; df=15; do a=5 to 2 by -1; q=probmcm("range",.,1-alpha,df,a); output; end; run; ods html; proc print data=pgm4_29 noobs;run; ods html close; quit;</pre> | |
| | 调用 probmcm 函数 |

结果显示为:

| alpha | df | a | q |
|-------|----|---|---------|
| 0.05 | 15 | 5 | 4.36699 |
| 0.05 | 15 | 4 | 4.07597 |
| 0.05 | 15 | 3 | 3.67338 |
| 0.05 | 15 | 2 | 3.01432 |

2.10 样本统计函数

2.10.1 样本统计函数简介

样本统计函数见表 2-11。

表 2-11 样本统计函数

| 函 数 | 描述 |
|--|----------------|
| CMISS(x, y, ...) | 计算缺失值的数量 |
| CSS(x, y, ...) | 计算校正平方和 |
| CV(x, y, ...) | 计算变异系数 |
| EUCLID(x, y, ...) | 返回非缺失值的欧氏范数 |
| GEOMEAN(x, y, ...) | 计算几何均值 |
| HARMEAN(x, y, ...) | 计算调和平均值 |
| IQR(x, y, ...) | 计算四分位数范围 |
| KURTOSIS(x, y, ...) | 计算峰度系数 |
| LARGEST(k, x, y, ...) | 计算第 k 个最大非缺失值 |
| MAD(x, y, ...) | 由中位数计算中位绝对变差 |
| MAX(x, y, ...) | 求最大值 |
| MEAN(x, y, ...) | 计算非缺失值的算术平均值 |
| MEDIAN(x, y, ...) | 求中位数 |
| MIN(x, y, ...) | 求最小值 |
| MISSING(num expression character expression) | 计算自变量的缺失数 |
| N(x, y, ...) | 计算样本个数, 不包括缺失值 |
| NMISS(x, y, ...) | 计算样本中缺失值的个数 |
| ORDINAL(count, x, y, ...) | 返回部分列表的最大值 |
| PCTL(percentage, x, y, ...) | 计算与百分比相对应的百分点 |
| RANGE(x, y, ...) | 计算极差 |
| RMS(x, y, ...) | 计算均方根 |
| SKEWNESS(x, y, ...) | 计算偏度系数 |
| SMALLEST(k, x, y, ...) | 计算第 k 个最小非缺失值 |
| STD(x, y, ...) | 计算标准差 |
| STDERR(x, y, ...) | 计算标准误 |
| SUM(x, y, ...) | 计算总和 |
| SUMABS(x, y, ...) | 计算非缺失值的绝对值的和 |
| USS(x, y, ...) | 计算平方和 |
| VAR(x, y, ...) | 计算方差 |

2.10.2 用 MEAN、MAX 与 MIN 函数计算算术均值、最大值与最小值

【例 2-141】 计算 234, -56, 9, 11 的算数均值、最大值与最小值。

具体 SAS 程序如下, 设程序名为 SASTJFX2_145.sas。

| 程序 | 说明 |
|--------------------------|------------|
| data SASTJFX2_145; | |
| x=mean(234, -56, 9, 11); | 调用 mean 函数 |
| y=max(234, -56, 9, 11); | 调用 max 函数 |
| z=min(234, -56, 9, 11); | 调用 min 函数 |
| put x = y = z =; | |
| run; | |

结果显示为：

x = 49.5 y = 234 z = -56, 这 4 个数的算数均值为 49.5, 最大值为 234, 最小值为 -56。

2.10.3 用 SUM、USS 与 CSS 函数计算和、未校正平方和与校正平方和

【例 2-142】 计算 157, 5, 88 的和、未校正平方和与校正平方和。

具体 SAS 程序如下, 设程序名为 SASTJFX2_146.sas。

| 程序 | 说明 |
|--------------------|-----------|
| data SASTJFX2_146; | |
| x = sum(157,5,88); | 调用 sum 函数 |
| y = uss(157,5,88); | 调用 uss 函数 |
| z = css(157,5,88); | 调用 css 函数 |
| put x = y = z =; | |
| run; | |

结果显示为：

x = 250 y = 32418 z = 11584.666667, 这 3 个数的和为 250, 未校正平方和为 32418, 校正平方和为 11584.666667。

2.10.4 用 VAR、STD、STDERR 和 CV 函数计算方差、标准差、标准误与变异系数

【例 2-143】 计算 512, 7, 80, 121 的方差、标准差、标准误与变异系数。

具体 SAS 程序如下, 设程序名为 SASTJFX2_147.sas。

| 程序 | 说明 |
|----------------------------|--------------|
| data SASTJFX2_147; | |
| x1 = var(512,7,80,121); | 调用 var 函数 |
| x2 = std(512,7,80,121); | 调用 std 函数 |
| x3 = stderr(512,7,80,121); | 调用 stderr 函数 |
| x4 = cv(512,7,80,121); | 调用 cv 函数 |
| put x1 = x2 = x3 = x4 =; | |
| run; | |

结果显示为：

x1 = 51211.333333 x2 = 226.29921196 x3 = 113.14960598 x4 = 125.72178442, 这 4 个数的方差为 51211.333333, 标准差为 226.29921196, 标准误为 113.14960598, 变异系数为 125.72178442。

2.10.5 用 SKEWNESS 和 KURTOSIS 函数计算偏度系数与峰度系数

【例 2-144】 计算 3,5,15,25,35 的偏度系数和峰度系数。

具体 SAS 程序如下, 设程序名为 SASTJFX2_148.sas。

| 程序 | 说明 |
|-----------------------------|----------------|
| data SASTJFX2_148; | |
| x = skewness(3,5,15,25,35); | 调用 skewness 函数 |
| y = kurtosis(3,5,15,25,35); | 调用 kurtosis 函数 |
| put x = y =; | |
| run; | |

结果显示为：

$x = 0.4622255068$ $y = -1.568741052$ ，这 5 个数的偏度系数为 0.4622255068，峰度系数为 -1.568741052。

2.10.6 用 NMISS 函数计算缺失值的个数

【例 2-145】 计算 1,0,-1,1,.,-1,0,1,1,.,1,0,-1,-1,0 的缺失值个数。
具体 SAS 程序如下，设程序名为 SASTJFX2_149.sas。

| 程序 | 说明 |
|--|-------------|
| data SASTJFX2_149; x=nmiss(1,0,-1,1,1,.,-1,0,1,1,.,1,0,-1,-1,0); put x=; run; | 调用 nmiss 函数 |

结果显示为：
 $x = 2$ ，有 2 个缺失值。

2.11 随机数函数

2.11.1 随机数函数简介

随机数函数见表 2-12。

表 2-12 随机数函数

| 函 数 | 描述 | 函 数 | 描述 |
|--------------------------------|---------------|-----------------------------|--------------|
| NORMAL(seed) | 计算服从正态分布的随机数 | RANNOR(seed) | 产生服从正态分布的随机数 |
| RANBIN(seed,n,p) | 产生服从二项式分布的随机数 | RANPOI(seed,m) | 产生服从泊松分布的随机数 |
| RANCAU(seed) | 产生服从柯西分布的随机数 | RANTBL(seed,p1,...pi,...pn) | 产生服从离散分布的随机数 |
| RAND('dist',parm-1,...,parm-k) | 根据特定的分布产生随机数 | RANTRI(seed,h) | 产生服从三角分布的随机数 |
| RANEXP(seed) | 产生服从指数分布的随机数 | RANUNI(seed) | 产生服从均匀分布的随机数 |
| RANGAM(seed,a) | 产生服从伽马分布的随机数 | UNIFORM(seed) | 产生服从均匀分布的随机数 |

2.11.2 用 NORMAL 函数或 RANNOR 函数产生正态分布的随机数

【例 2-146】 用 NORMAL 函数产生服从 $N(0,1)$ 的正态分布随机数。
具体 SAS 程序如下，设程序名为 SASTJFX2_150.sas。

| 程序 | 说明 |
|--|--|
| data SASTJFX2_150; retain _seed_0; m=0; s=1; do _i_=1 to 1000; x=m+s* normal(_seed_); output; end; drop _seed__i_; run; | 将 seed 赋值为 0 定义平均数 m 为 0 定义标准差 s 为 1 调用 normal 函数 |

结果产生 1000 个服从 $N(0,1)$ 的正态分布随机数，输出名称为 SASTJFX2_150 临时数据集。

【例 2-147】 用 RANNOR 函数产生服从 $N(0,1)$ 的正态分布随机数。
具体 SAS 程序如下，设程序名为 SASTJFX2_151.sas。

| 程序 | 说明 |
|--|--|
| <pre>data SASTJFX2_151; retain _seed_0; m=0; s=1; do _i_=1 to 1000; x=m+sqrt(s)* rannor(_seed_); output; end; drop _seed__i_; run;</pre> | 将 seed 赋值为 0 定义平均数 m 为 0 定义标准差 s 为 1 调用 rannor 函数 |

结果产生 1000 个服从 $N(0,1)$ 的正态分布随机数，输出名称为 SASTJFX2_151 临时数据集。
NORMAL 函数和 RANNOR 函数的区别如下。

NORMAL(seed) 函数: seed 为整数变量，产生随机数的种子，seed 为 0，或 5 位、6 位、7 位的奇数。RANNOR(seed) 函数: seed 取值范围为 $1 \sim 2^{31} - 1$ ，如果 seed 小于 0，则采用当前时间作为 seed 产生随机数。正态分布随机数函数 NORMAL 和 RANNOR 是同质的，但 NORMAL 没有相对应的 CALL 子程序。

2.11.3 用 UNIFORM 或 RANUNI 函数产生均匀分布的随机数

【例 2-148】 用 UNIFORM 函数产生在区间 $[-5,1]$ 上的均匀分布随机数。
具体 SAS 程序如下，设程序名为 SASTJFX2_152.sas。

| 程序 | 说明 |
|--|--|
| <pre>data SASTJFX2_152; retain _seed_0; a=-5; b=1; do _i_=1 to 1000; x=a+(b-a)* uniform(_seed_); output; end; drop _seed__i_; run;</pre> | 将 seed 赋值为 0 区间下限为 -5 区间上限为 1 调用 uniform 函数 |

结果产生 1000 个在区间 $[-5,1]$ 上的均匀分布随机数，输出名称为 SASTJFX2_152 临时数据集。

【例 2-149】 用 RANUNI 函数产生在区间 $[-5,1]$ 上的均匀分布随机数。
具体 SAS 程序如下，设程序名为 SASTJFX2_153.sas。

| 程序 | 说明 |
|---|---|
| <pre>data SASTJFX2_153; retain _seed_0; a=-5; b=1; do _i_=1 to 1000; x=a+(b-a)* ranuni(_seed_); output; end; drop _seed__i_; drop _seed__i_; run;</pre> | 将 seed 赋值为 0 区间下限为 -5 区间上限为 1 调用 ranuni 函数 |

结果产生 1000 个在区间 $[-5, 1]$ 上的均匀分布随机数, 输出名称为 SASTJFX2_153 临时数据集。

UNIFORM 和 RANUNI 函数的区别如下。

UNIFORM(seed) 函数: seed 必须是常数, 为 0, 或 5 位、6 位、7 位的奇数。RANUNI(seed) 函数: seed 为小于 $2^{31} - 1$ 的任意常数。均匀分布函数 UNIFORM 和 RANUNI 是同质的, 但 UNIFORM 没有相对应的 CALL 子程序。

2.11.4 用 RANEXP 函数产生指数分布的随机数

【例 2-150】 产生参数为 1.3 的指数分布的随机数。

具体 SAS 程序如下, 设程序名为 SASTJFX2_154.sas。

| 程序 | 说明 |
|--|---|
| data SASTJFX2_154; retain _seed_ 0; lambda=1.3; do _i_=1 to 1000; x=ranexp(_seed_)/lambda; output; end; drop _seed_ _i_ lambda; run; | 将 seed 赋值为 0 参数为 1.3 调用 ranexp 函数 |

结果产生 1000 个参数为 1.3 的指数分布随机数, 输出名称为 SASTJFX2_154 临时数据集。

2.11.5 用 RANBIN 函数产生二项分布的随机数

【例 2-151】 产生符合二项分布(10, 0.5)的随机数。

具体 SAS 程序如下, 设程序名为 SASTJFX2_155.sas。

| 程序 | 说明 |
|--|--------------|
| data SASTJFX2_155; retain _seed_ 0; n=10; p=0.5; do _i_=1 to 1000; x=ranbin(_seed_, n, p); output; end; drop _seed_ _i_; run; | 调用 ranbin 函数 |

结果产生 1000 个符合二项分布(10, 0.5)的随机数, 输出名称为 SASTJFX2_155 临时数据集。

2.11.6 用 RANPOI 函数产生泊松分布的随机数

【例 2-152】 产生参数为 3.7 的泊松分布的随机数。

具体 SAS 程序如下, 设程序名为 SASTJFX2_156.sas。

| 程序 | 说明 |
|--|--------------|
| <pre>data SASTJFX2_156; retain _seed_0; lambda=3.7; do _i_=1 to 1000; x=ranpoi(_seed_,lambda); output; end; drop _seed_ _i_ lambda; run;</pre> | 调用 ranpoi 函数 |

结果产生 1000 个参数为 3.7 的泊松分布随机数，输出名称为 SASTJFX2_156 临时数据集。

值得注意的是：在同一个数据步中对同一个随机数函数的多次调用将得到不同的结果，但不同数据步中从同一种子出发将得到相同的随机数序列。随机数种子如果取 0 或者负数则种子采用系统日期时间。

2.12 SAS call 子程序

SAS 系统提供了一系列 CALL 子程序，用于产生随机数或执行其他系统功能。所有 SAS CALL 子程序均由 CALL 语句激活，即函数的名字必须写在 CALL 语句的关键字 CALL 后面。

2.12.1 随机数子程序

随机数子程序主要有 9 个，见表 2-13。

表 2-13 随机数子程序

| 随机数子程序 | 描 述 | 随机数子程序 | 描 述 |
|--------------------------|----------------|---|---------------|
| CALL RANBIN(seed,n,p,x) | 返回来自二项式分布的随机变量 | CALL RANPOI(seed,m,x) | 返回来自泊松分布的随机变量 |
| CALL RANCAU(seed,x) | 返回来自柯西分布的随机变量 | CALL RANTBL(seed,p1, . . . pi, . . . pn,x) | 返回来自离散分布的随机变量 |
| CALL RANEXP(seed,x) | 返回来自指数分布的随机变量 | CALL RANTRI(seed,h,x) | 返回来自三角分布的随机变量 |
| CALL RANGAM(seed,a,x) | 返回来自伽马分布的随机变量 | CALL RANUNI(seed,x) | 返回来自均匀分布的随机变量 |
| CALL RANNOR(seed,x) | 返回来自正态分布的随机变量 | | |

2.12.2 其他子程序

这类子程序 SAS 提供了很多，较为常用的有以下 5 种，见表 2-14。

表 2-14 其他子程序

| 其他子程序 | 描 述 |
|----------------------|--------------------------|
| CALL LABEL(v1,v2) | 指定字符变量 v2 的值为变量 v1 的标签 |
| CALL SYMPUT(x1,x2) | 通过 DATA 步对宏变量提供信息 |
| CALL SYSTEM(command) | 发布主机系统命令 |
| CALL VNAME(v1,v2) | 指定变量 v1 的名字作为变量 v2 的值 |
| CALL EXECUTE(s) | 规定一个字符表达式来得到宏调用或者 SAS 语句 |

2.12.3 随机数子程序的运用

SAS 系统提供一系列随机数子程序，使得用户可以比用随机数函数更好地控制种子流和随机数流。除 NORMAL 和 UNIFORM 这两个函数之外，所有随机数函数都有一个相应的子程序。

随机数发生器子程序用 CALL 语句产生。CALL 语句的一般格式为：

```
CALL routine(seed, <argument, >variate);
```


其中, routine 是某个 SAS 随机数函数的名字; seed 是存放当前种子值的变量名字。variate 是存放生成的随机数的变量名字; argument 是特殊分布要求的参数。seed 变量在 CALL 语句第一次执行前被赋予初值, 例如, 用赋值语句或 RETAIN 语句赋值。

CALL 语句执行之后, seed 保存这个流的当前种子值, variate 保存生成的随机数。

当用 DATA 步来生成几个随机数流时, 使用 CALL 子程序比使用随机数函数效果更好。虽然用一些随机数函数也能创建几个随机数集, 但它们都属于一个流。

例如:

```
data SASTJFX2_157;
  retain seed1 seed2 123456789;
  do i=1 to 5;
    x1 = ranuni(seed1);
    x2 = ranuni(seed2);
    output;
  end;
ods html;
proc print;
  title '使用随机数函数';
run;
ods html close;
```

输出结果如下:

| | | 使用随机数函数 | | | |
|-----|-----------|-----------|---|---------|---------|
| Obs | seed1 | seed2 | i | x1 | x2 |
| 1 | 123456789 | 123456789 | 1 | 0.95542 | 0.87942 |
| 2 | 123456789 | 123456789 | 2 | 0.99394 | 0.82913 |
| 3 | 123456789 | 123456789 | 3 | 0.99744 | 0.25082 |
| 4 | 123456789 | 123456789 | 4 | 0.21090 | 0.83717 |
| 5 | 123456789 | 123456789 | 5 | 0.58702 | 0.71544 |

种子 seed1 和 seed2 初始值是一样的, 但是生成并存放在 X1 和 X2 的随机数是不相同的。这是因为在 DATA 步只产生一个随机数流, X1 和 X2 中的随机数都是 seed1 产生的结果, 当产生 X1 第 1 个值后, 再产生 X2 第 1 个值, 再产生 X1 第 2 个值, 如此继续。seed2 在这里没有起到任何作用。

如果使用 CALL 子程序, 就可以产生多个随机数流。

例如:

```
data SASTJFX2_158;
  retain seed3 123456789;
  retain seed4 123456789;
  retain seed5 157903284;
  do i=1 to 5;
    call ranuni(seed3,x3);
    call ranuni(seed4,x4);
    call ranuni(seed5,x5);
    output;
  end;
ods html;
proc print;
  title '调用随机数 CALL 子程序';
run;
ods html close;
```

输出结果如下：

| 调用随机数 CALL 子程序 | | | | | | | |
|----------------|------------|------------|------------|---|---------|---------|---------|
| Obs | seed3 | seed4 | seed5 | i | x3 | x4 | x5 |
| 1 | 2051752218 | 2051752218 | 412284237 | 1 | 0.95542 | 0.95542 | 0.19198 |
| 2 | 1888541278 | 1888541278 | 2088756581 | 2 | 0.87942 | 0.87942 | 0.97265 |
| 3 | 2134478923 | 2134478923 | 2049836366 | 3 | 0.99394 | 0.99394 | 0.95453 |
| 4 | 1780549980 | 1780549980 | 2083277757 | 4 | 0.82913 | 0.82913 | 0.97010 |
| 5 | 2141992067 | 2141992067 | 1262476829 | 5 | 0.99744 | 0.99744 | 0.58789 |

这里运用随机数 CALL 子程序产生了三个独立的种子流和随机数流，因为 seed3 = seed4，所以生成的并存放在 X3 和 X4 中的随机数是完全相同的，seed3、seed4 与 seed1 的初始值是相同的，所以 X3 中的随机数即是 X1 和 X2 中随机数共同组成，X4 同理。seed5 被赋予了不同的初始值，因此产生了不同的随机数流。使用 CALL 语句用户随时可以查看到当前的种子值，而使用随机数函数用户只能看到初始值，如 seed1 中 5 个观测只能看到 seed1 被赋予的初始值 123456789。

(王琪 郭春雪 孙日扬)

第3章 SAS 高级编程技术介绍

在运用 SAS 软件实现统计分析时，若主要利用其过程步，则数据步中一般只需要运用简单的 SAS 语句、SAS 函数就可达到目的。然而，当用户找不到合适的 SAS 过程时，可能就需要利用 SAS 中提供的高级编程技术了。它们包括 SAS ODS(Output Delivery System, 输出传送系统, 以一个子系统的面貌出现)、SAS 宏、SAS/SQL(Structured Query Language, 结构查询语言, 以一个 SAS 过程的面貌出现)、SAS 数组和 SAS/IML(Interactive Matrix Language, 交互式矩阵语言, 以一个 SAS 产品或模块的面貌出现)。本章将扼要介绍以上五大类 SAS 高级编程技术。

3.1 SAS ODS 介绍

3.1.1 概述

传统的 SAS 过程输出结果均是为传统的针式打印机而设计的，这种方式存在着一定的局限性：

- 传统的 SAS 输出使用的是等宽的列表字体，而在各种文件编辑系统下，其使用往往会受到限制。
- 一些常用过程通常只产生输出结果，并不创建数据集，如果我们能将结果转化为输出数据集，那么我们就可以将其作为其他过程或者数据步的输入数据集，提高了程序的连贯性。

ODS(Output Delivery System) 的出现就是为突破这些局限性的，使输出结果个性化更加容易。用户在利用 ODS 进行生成、存储以及复制 SAS 输出结果时有很大的自由发挥空间，丰富的格式选项有助于用户有效地定制 SAS 输出，更充分地利用分析结果。

3.1.2 ODS 特点和常用输出目标

1. ODS 特点

- ODS 将未加工的数据与表定义结合起来，并生成一个或多个输出对象，这些目标可以被传递到所有的 ODS 目标。用户通过选择 ODS 目标来控制输出形式，当前可用的 ODS 目标能够生成如下几种输出形式：
 - 传统的等宽字体的 SAS 输出。
 - 输出 SAS 数据集。
 - 输出包含多个输出对象树状结构的 ODS document 文档。
 - 输出为 PostScript 或 PDF 文件格式用于高分辨率打印。
 - 输出多样的格式化标记语言，如 HTML。
 - 输出为 RTF 文件格式用于 MS word 软件打开和编辑。
- ODS 利用表定义来明确 SAS 过程步和数据步的输出结构，用户可以通过修改这些定义来输出想要的结构。

- ODS 提供了一种将输出对象传送到 ODS 目标的方式。例如，UNIVARIATE 过程生成了五个输出对象，我们很容易将这些对象输出成 HTML 文档、SAS 数据集、传统 SAS 表格输出或者高清晰打印文件。不同的输出对象可以对应不同的 ODS 目标。
- 在 SAS 环境下，ODS 在结果文件夹中存储了每一个输出对象的连接。
- 由于格式操作都集中在 ODS 内部进行，故一个新加入的 ODS 目标不会影响到任何过程步或者数据步的正常运行。数据步和所有支持 ODS 的过程步都可以调用这个新 ODS 目标。
- 利用 ODS，只需运行过程一次，我们就可以从一个数据源获取各种 ODS 目标生成的输出，这样既节省了时间，又节省了系统资源。

2. ODS 目标

ODS 目标主要有以下两类。

(1) SAS 格式目标：生成一个由 SAS 控制和解释的输出，如 SAS 数据集、SAS 输出列表，以及 SAS ODS 文件。

(2) 第三方格式目标：生成一个可以应用样式、标记语言，以及用于打印的输出，如生成 PostScript、HTML、XML，或者是用户自行创建的其他格式的文档。

• DOCUMENT 目标

通过 DOCUMENT 目标可以用不同方式对数据进行重建和操纵，可以不用重新运行程序而将数据传递到其他目标。DOCUMENT 目标的数据流都是未加工的形式，用户可以根据需求进行个性化。DOCUMENT 目标提供 GUI(图形用户界面)接口，从而使许多操作可以直接通过 GUI 完成。当然，通过 ODS DOCUMENT 语句和 DOCUMENT 过程辅助各种指令一样可以完成相同的操作。

在 SAS 9 之前，过程步和数据步生成的输出都发送到用户指定的目标，并且可以同时指定多个目标，但如果用户试图将输出发送到一个事先未指定的目标，则需要重新运行程序。DOCUMENT 目标则省略了这一重复操作，它可以存储输出主体并将结果再现给其他目标。

• LISTING 目标

LISTING 目标产生的输出和传统 SAS 输出基本类似，它是 SAS 初始运行时默认的输出目标，也就是说，我们时刻都在使用 ODS。

由于大多数过程都共享着部分相同的表定义，所以不同过程的输出也有着一定程度的一致性。例如，不同的过程会以相同的方式生成 ANOVA 表，因为它们使用相同的模板描述这个表。然而，有三个过程不使用默认表定义产生输出，分别是 PRINT 过程、REPORT 过程和 TABULATE 过程，这类过程使用的表格结构通常需要用户在代码中指定。

• OUTPUT 目标

OUTPUT 目标可以产生 SAS 输出数据集，便于用户做进一步分析，有时将不同数据集中相同的统计量输出到一个报告中也可以使用 OUTPUT 目标。将运行结果转化为输出数据集之后，就可以运用各种数据步指令进行数据集操作，其使用方法与处理其他 SAS 数据集时没有差别。

• HTML 目标

HTML 目标默认生成的是 HTML4.0 文档，使用 HTML3 语句可以生成 HTML3.2 文档，文档中包含指定过程的输出结果。

ODS HTML 语句提示输出 HTML 文档，这种网页格式包含丰富的框架和样式表，在输出量

较大时便于浏览和查阅。HTML 目标适用于需要在线使用结果的场合，对于需要打印的情形请使用 PRINTER 目标。

- MARKUP 目标

就像表定义描述表的设计，样式属性描述输出样式，TAGSET 用于描述怎样产生标记语言输出。用户既可以使用 SAS 支持的外部 TAGSET，也可以利用 TEMPLATE 过程创建自己的 TAGSET，这使得 MARKUP 目标的应用十分灵活。

- PRINTER 目标

PRINTER 目标主要是产生用于物理打印的等便携式的格式文件。这种文件通常都遵循严格的页面规则，也就是说，用户不能编辑或更改这些格式的文件内容。因此，PRINTER 目标产生的输出一般都被作为报告的最终输出形式。

- RTF 目标

RTF 目标的输出主要面向 MS Word，利用其他软件对 RTF 格式文件进行操作可能会出错。

由于 ODS 本身没有对页面进行任何垂直测量，这将导致表格出现的位置不一定是最理想的，我们并不希望一个表格在中间位置（除非是用户指定的位置）被分开，而是希望输出中尽量保证每个表格的完整性。MS Word 是需要知道表格的宽度，并且在表格超过页面宽度时不能自动调整，而 ODS 可以限制文本内容的水平宽度，过宽的表格会被分解成多个表单元。

简而言之，当使用 RTF 目标时，ODS 控制文本内容的水平宽度，MS Word 控制文本内容的垂直宽度，因为 MS Word 可以决定一个页面的空间大小。

3.1.3 常用 ODS 语句

1. PUT 语句

ODS 利用 PUT 语句通过一个特殊的缓冲期为输出数据添加各种各样的格式，这个语句只在数据步中有效，并且是可执行的。本节只介绍 ODS 形式的 PUT 语句用法，该语句的使用格式如下：

```
PUT <specification> <_ODS_> <@ |@@>;
Specification: 指定要输出的变量内容和变量位置;
```

ODS: 将变量值写到输出窗口中。使用该选项时，必须在之前使用 FILE PRINT ODS COLUMNS 语句。如果没有指定 COLUMNS，则变量的输出顺序将与缓冲器中的变量顺序相同。

@ |@@: 在 DATA 步不断循环中，保持输出的观测行固定不变。

ODS 列指针与不使用 ODS 的 PUT 语句列指针有细微差别。ODS 列指针不是指单个字符的输出位置，而是指给定变量取值的输出位置，因此，移动指针即从一个列变量位置移动到另一个列变量位置。COLUMN1 表示指定变量输出在第一列的位置，COLUMN2 表示指定变量输出在第二列的位置，以此类推。列指针的一般用法是“@ ODS-COLUMN”，这个数值必须是正的；“+ ODS-COLUMN”表示将指针移动指定的行数，当这个值为正时，指针向右移动；反之，向左移动，为 0 时，指针不移动；“@ 'COLUMN-NAME'”表示移动到指定名字的变量。

ODS 行指针与普通行指针用法基本相同。“#LINE”表示将指针移动到指定行；“/”表示把指针移动到下一行的第一列。

【例 3-1】 利用 PUT 语句输出程序中的数据。设程序名为 SASTJFX3_1.sas。

```
data _null_;
input id name $ score@@ ;
FILE PRINT ODS;
put @ 2 name + (-2) id +1 score;
cards;
1 ZS 94 2 FQ
3 NQ 99 4 DZ
5 LR 95 6 NS
7 SS 97 8 FS
9 DK 98 10 ND
;
run;
quit;
```

运行结果：

| id | name | score |
|----|------|-------|
| 1 | ZS | 94 |
| 2 | FQ | 98 |
| 3 | NQ | 99 |
| 4 | DZ | 96 |
| 5 | LR | 95 |
| 6 | NS | 99 |
| 7 | SS | 97 |
| 8 | FS | 98 |
| 9 | DK | 98 |
| 10 | ND | 97 |

这里，我们关注一下 ODS 列指针的用法，“@ 2 NAME”表示在第二列输出 NAME 变量的全部取值，data 步循环每次在某一列输出完变量值之后，列指针都会自动移动到下一列，接下来的“+ (-2)id”就是将列指针从第三列左移两列到第一列，并写下变量 ID 的取值，程序这样执行之后，列指针应指向第二列，那么只需要向右移动一列，列指针即指向第三列，然后写下变量 SCORE 的取值。

将程序中对应语句做如下替换：

```
FILE PRINT ODS = (variables = (score name));
put_ods_;
```

运行结果：

| score | name |
|-------|------|
| 94 | ZS |
| 98 | FQ |
| ... | ... |

这里，我们利用 VARIABLES 指定了要输出的变量名称和输出顺序，PUT 语句就会按规定将对应的变量取值全部输出出来。

【例 3-2】 PUT 语句中 COLUMNS 选项的使用。设程序名为 SASTJFX3_2.sas。

```

proc template;
  define table scorelist;
    column N S;
  end;
run;
data _null_;
  input id name $ score @@ ;
  FILE PRINT
  ODS = (template = 'scorelist'
  columns = (S = score N = name));
  put ods_;
  cards;
  1 ZS 94 2 FQ 98
  3 NQ 99 4 DZ 96
  5 LR 95 6 NS 99
  7 SS 97 8 FS 98
  9 DK 98 10 ND 97
;
run;

```

运行结果：

| name | score |
|------|-------|
| ZS | 94 |
| FQ | 98 |
| NQ | 99 |
| DZ | 96 |
| LR | 95 |
| NS | 99 |
| SS | 97 |
| FS | 98 |
| DK | 98 |
| ND | 97 |

关于 TEMPLATE 过程的用法将在后面加以介绍，这里只强调 COLUMNS 选项的用法。应用该选项，必须先通过 TEMPLATE 过程定义一个表名称，再定义列变量名称，然后在 COLUMNS 中指定每一列所对应的输入数据中的变量。注意到，输出变量顺序与 TEMPLATE 过程的列变量名称定义顺序是相同的。

2. ODS TRACE 语句

该语句使用格式如下：

```

ODS TRACE ON </option(s) >;
ODS TRACE OFF;

```

该语句主要控制是否在 SAS 日志中给出一个对象的跟踪记录，默认状态是不给出。

常用选项包括：EXCLUDED，令跟踪记录包含被剔除输出对象的相关信息；LABEL，在跟踪记录中给出输出对象的路径标签；LISTING，在 LISTING 目标中输出跟踪记录，即每个输出对象前面都紧跟着描述这个对象的跟踪信息。

ODS 生成的输出对象是将数据与数据成分通过表定义组合起来。跟踪记录提供了以下关于数据成分、表定义、输出对象的主要相关信息。

NAME：输出对象的名称，可以用该名称来引用这个输出对象。

LABEL: 输出对象的标签, 简略地描述输出对象的内容。
DATA NAME: 创建这个输出对象的数据成分的名称。仅当与输出对象名称不同时给出。
DATA LABEL: 描述数据的内容。
TEMPLATE: ODS 给输出对象规定格式所使用的表定义的名称。
PATH: 输出对象的路径, 可以使用该路径来引用这个输出对象。
对于 SAS 程序生成的众多输出对象, 我们要知道怎样保留有用的对象, 剔除不需要的对象。下面介绍几种指定输出对象的方法。

- 全路径: 如“Univariate. City_Pop_90. TestsForLocation”就是一个全路径。
- 部分路径: 是对应全路径的一部分, 可以是任意“.”后面的部分, 以前面的全路径为例, 它的部分路径可以是“City_Pop_90. TestsForLocation”或者“TestsForLocation”。
- 带双引号的标签: 如“The UNIVARIATE Procedure”。
- 标签路径: 如果设定了 LABEL 选项, 跟踪记录中会给出标签路径, 如“"The UNIVARIATE Procedure". "CityPop_90". "Tests For Location"”。

部分标签路径: 与部分路径类似, 以上面的标签路径为例, 它的部分标签路径可以是“"CityPop_90". "Tests For Location"”或者是“"Tests For Location"”。

【例 3-3】 输出 SAS 程序的跟踪记录。设程序名为 SASTJFX3_3. sas。

```
ods trace on/label listing;
proc univariate data=sashelp.class;
class sex;
var height;
run;
ods trace off;
quit;
```

部分运行结果:

Output Added:

名称: Moments
标签: 矩
模板: base.univariate.Moments
路径: Univariate.Height.男.Moments
标签路径: 'Univariate PROCEDURE'. 'Height'. 'Sex = 男' '矩'

| 矩 | | | |
|-------|------------|--------|------------|
| N | 10 | 权重总和 | 10 |
| 均值 | 63.91 | 观测总和 | 639.1 |
| 标准差 | 4.93793704 | 方差 | 24.3832222 |
| 偏度 | 0.04095917 | 峰度 | -0.934876 |
| 未校平方和 | 41064.33 | 校正平方和 | 219.449 |
| 变异系数 | 7.72639187 | 标准误差均值 | 1.5615128 |

这里只给出了部分运行结果, 程序中 ODS TRACE 语句中设定了 LABEL 和 LISTING 选项, 因此跟踪记录中会给出标签路径, 并且跟踪记录会在输出窗口中给出。如果去掉 LISTING 选项, 跟踪记录会在 SAS 日志中输出。例子是在 SAS 中文版上运行的程序, 可以看到跟踪记录包括输出对象名称 (NAME)、标签 (LABEL)、模板 (TEMPLATE)、路径 (PATH)、标签路径

(LABEL PATH), 后面分别紧跟着相应的内容, 跟踪记录之后是一个完整的输出对象, 通过利用路径或标签等信息就可以特别指定该输出对象。

3. 常用控制语句

• ODS EXCLUDE 语句

该语句使用格式如下:

```
ODS <ODS-destination> EXCLUDE exclusion(s) [ALL | NONE];
```

exclusion(s): 指定一个或多个要剔除的输出对象。这里详细介绍怎样指定输出对象。首先, 必须知道将要运行的 SAS 程序都会产生哪些输出对象, 然后要了解这些对象在 SAS 中怎样去描述。上节讲到的 ODS TRACE 语句可以把与输出对象有关的各种信息写到 SAS 日志中, 而这些信息可以成为描述该对象的有力工具。对几种指定输出对象的方法已在前面介绍过, 如全路径、部分路径、标签路径等。

ALL: 剔除所有的输出对象。

NONE: 不剔除所有的输出对象。

ODS-destination: 指定要在哪个 ODS 目标中剔除某输出对象。

WHERE = where-expression: 利用 WHERE 语句可以剔除满足指定条件的输出对象。如“ods exclude where = (_name_? 'Histogram');”表示剔除名称中含有“HISTOGRAM”的输出对象。WHERE 表达式可以是算术或者逻辑表达式, 是由一系列操作符和操作对象构成的。“_NAME_”是 SAS 自定义的一个特殊操作对象, 用以指代输出对象的名称, 类似的还有“_LABEL_”指代输出对象的标签, “_LABELPATH_”指代输出对象的标签路径, “_PATH_”指代输出对象的全路径或部分路径。关于操作符的种类及含义在其他章节已经详细介绍过, 这里不再赘述。

【例 3-4】 不同输出目标剔除不同输出对象。设程序名为 SASTJFX3_4.sas。

```
ods html;
ods html exclude where = (_path_?"Height");
ods listing exclude where = (_path_?"Weight");
proc univariate data = sashelp.class;
var height weight;
run;
ods html close;
quit;
```

部分运行结果如图 3-1 所示。



图 3-1 SAS 呈现输出结果的树状图之一

例 3-4 中分别对两个输出目标进行了操作：一个是 LISTING 目标，另一个是 HTML 目标。我们将 HTML 目标中路径包含“HEIGHT”的输出对象全部剔除掉，对应地将 LISTING 目标中路径包含“WEIGHT”的输出对象全部剔除掉，就得到了上面的运行结果。可以看到，LISTING 目标中只含有“HEIGHT”的相关输出信息，而 HTML 目标中只含有“WEIGHT”的相关输出信息。

• ODS SELECT 语句

该语句使用格式如下：

```
ODS <ODS-destination> SELECT exclusion(s) [ALL | NONE];
```

其中，各参数含义和该语句用法都可参照 ODS EXCLUDE 部分讲述的内容，主要区别在于对输出对象的剔除改为对输出对象的选择，最后 ODS 目标只会给出被选择的输出对象。

【例 3-5】 不同输出目标选择不同输出对象。设程序名为 SASTJFX3_5.sas。

将例 3-4 中的

```
ods html exclude where = (_path_?"Height");
ods listing exclude where = (_path_?"Weight");
```

替换成

```
ods html select where = (_path_?"Height");
ods listing select where = (_path_?"Weight");
```

部分运行结果如图 3-2 所示。



图 3-2 SAS 呈现输出结果的树状图之二

可以看到，得到的运行结果与上一例正好相反，如果我们去掉 ODS 后面的目标参数，那么该 SELECT 操作对所有输出目标都生效。读者也可以利用其他的输出对象引用方式，以 7.5 为例，挑选了两个变量的“MOMENTS”输出对象，即变量 HEIGHT 的“BasicMeasures”输出对象，变量 WEIGHT 的“TestsForLocation”输出对象：

```
ods listing select
Moments/* 名称* /
Univariate.Height.BasicMeasures /* 路径* /
'Univariate PROCEDURE'. 'Weight'. '位置检验'; /* 标签路径* /
```

• ODS SHOW 语句

该语句使用格式如下：

```
ODS <ODS-destination> SHOW;
```

ODS SHOW 语句主要用来决定将哪个输出目标的选择和排除列表写到 SAS 日志中去,省略该语句时,默认是输出总体的选择列表和排除列表。

【例 3-6】 利用 ODS SHOW 语句获得输出目标的相关列表。设程序名为 SASTJFX3_6.sas。

```
ods html;
ods show;
ods select Univariate.Weight.BasicMeasures;
ods html select where = (_path_?"Height" and
_name_ = 'BasicMeasures');
ods show;
ods listing exclude Moments
Univariate.Height.BasicMeasures
'Univariate PROCEDURE'. 'Weight'. '位置检验';
ods html show;
ods listing show;
proc univariate data = sashelp.class;
var height weight;
run;
ods html close;
quit;
```

日志窗口运行信息:

```
256 ods show;当前 OVERALL select 列表是: ALL
259 ods show;当前 OVERALL select 列表是:
1. Univariate.Weight.BasicMeasures
262 ods html show;当前 HTML select 列表是:
1. where = (_path_contains 'Height' and(_name_ = 'BasicMeasures'))
263 ods listing show;当前 LISTING exclude 列表是:
1. Moments2. Univariate.Height.BasicMeasures
3. 'Univariate PROCEDURE'. 'Weight'. '位置检验'
```

结果修改了格式并删除了部分行,以节省篇幅,信息中 256、259 等数字与例子无关,是提交的程序总行数。可以看到,两个不同输出目标之间的选择列表和排除列表都相互独立,且两种列表各自也互相独立,而 OVERALL 的选择列表和排除列表会影响到所有目标对应的列表,但对同一列表的最新一次操作会覆盖之前所有的操作。第一个 ODS SHOW 语句给出的是 OVERALL 选择列表,此时没有做任何输出对象选择操作,默认是 ALL;随后的 ODS SELECT 语句没有指定目标,是对所有已开启的输出目标的选择列表进行操作,紧接着的 ODS HTML SELECT 语句仅对 HTML 目标的选择列表有效,所以第二个 ODS SHOW 语句显示 OVERALL 选择列表仅包含 WEIGHT 的 BASICMEASURES,而没有包含变量 HEIGHT 的 BASICMEASURES;对于后面的两个 ODS SHOW 语句,读者可以自行解读。最后的结果应为每个输出目标选择列表中的输出对象,即在 HTML 目标中输出变量 HEIGHT 的 BASICMEASURES,在 LISTING 目标中输出 WEIGHT 的 BASICMEASURES。

4. ODS LISTING 语句

该语句控制 LISTING 目标的打开、管理和关闭等操作,该输出目标是 SAS 默认打开的。该语句使用格式如下:

```
ODS LISTING <action>;
ODS LISTING <DATAPANEL = number |DATA |PAGE > <FILE = file-specification>;
```

ACTION 主要包括 CLOSE、EXCLUDE、SELECT、SHOW 这四个选项，分别是关闭输出目标、将输出对象加入剔除列表、将输出对象加入选择列表、在 SAS 日志中显示当前剔除列表和选择列表内容。它们的具体功能请参考本章相关内容。

“FILE =”选项用来指定存储输出结果的文件，等号右边可以是一个外部文件名，也可以是一个指定外部文件的文件引用。如果忽略此选项，ODS 会把结果直接输出到 OUTPUT 窗口中。

LISTING 目标是在 SAS 版本 7 开发 ODS 之前一直使用的传统输出目标，也是被大多数用户所熟悉的。LISTING 目标常用来临时观测数据的分析结果，程序上不需做任何修改，也不需要设定繁杂的选项，因此执行起来简单而高效。但 LISTING 目标得到的结果不适宜保存、修改、传阅以及打印，所以，对于正规的实验数数据分析或者调查设计研究等方面，我们鼓励使用 HTML、OUTPUT、RTF、PDF 等目标来输出结果。

5. 常用第三方格式输出目标语句

• ODS HTML 语句

该语句控制 HTML 目标的打开、管理和关闭等操作，生成 HTML4.0 文档来存储输出结果，属于第三方文件格式类。当打开 HTML 目标时，SAS 会自动创建一个与主体文件关联的样式表，用以限定 HTML 文档的显示格式，我们在 SAS 程序内无法对这个样式表做任何修改，即“STYLE =”选项不起任何作用，读者需要直接修改样式表来更改文档内容的显示格式。该语句使用格式如下：

```
ODS HTML < (<ID = >identifier) > <action>;  
ODS HTML < (<ID = >identifier) > <option(s)>;
```

如果不填写 ACTION 和 OPTIONS，ODS 会默认打开 HTML 目标并创建 HTML 输出。

ACTION 主要包括 CLOSE、EXCLUDE、SELECT、SHOW 这四个选项，具体功能请参考本章相关内容。

ODS HTML 为用户提供了很多选项，便于用户根据需求设计出想要的网页显示结果。表 3-1 总结了一些常用选项及其功能。

表 3-1 ODS HTML 常用选项及其功能

| 选 项 | 功 能 |
|-----------------|------------------------------|
| BASE = | 指定 HTML 文档链接和引用的 URL 起始部分 |
| BODY = | 指定包含 ODS 基本输出结果主体的文件 |
| CODE = | 指定包含有关样式信息的文件 |
| CONTENTS = | 指定包含 ODS 输出结果目录的文件 |
| CSSSTYLE = | 指定作用于结果的 CSS 样式表 |
| FRAME = | 指定将主体文件、目录文件、页码文件连接起来的总体框架文件 |
| GPATH = | 指定所有生成的图形输出的位置 |
| GTITLE NOGTITLE | 控制图形输出时标题的打印位置 |
| ID = | 同时运行同一输出目标的不同选项设定 |
| NEWFILE = | 在指定的起始点创建新的主体文件 |
| PAGE = | 指定包含每页主体文件的描述以及指向主体文件链接的页码文件 |
| PATH = | 指定外部文件或者 SAS HTML 文件目录的位置 |
| STYLE = | 指定用于输出文件的样式定义 |
| STYLESHEET = | 指定包含输出结果样式信息的文件 |
| TEXT = | 指定文档中插入的文本 |
| TITLE = | 指定位于浏览器窗口标题栏的文本 |

ODS HTML 语句是 ODS 标记语言家族的一员,同一类的还有 XML(Extensible Markup Language)、LaTeX 等。对于现在这个网络流行的时代,HTML 作为储存结果的文件格式显然有着广泛的适用性。

【例 3-7】 利用 ODS HTML 语句创建完整的逻辑网页。设程序名为 SASTJFX3_7.sas。

```
ods html
body = 'class_body.html' (title = 'CLASS_BODY')
contents = 'class_contents.html' (title = 'CLASS_CONTENTS')
frame = 'class_frame.html' (title = 'CLASS_FRAME')
page = 'class_page.html' (title = 'CLASS_PAGE')
path = 'D:\';
proc univariate data = sashelp.class;
var height weight;
run;
ods html close;quit;
```

运行结果如图 3-3 所示。



图 3-3 UNIVARIATE 过程输出结果的开始部分之表现一

上面展示的是一个完整的网页输出实例,程序中 ODS HTML 语句设定了诸多选项,BODY 选项设定了主体文件名,随后的 TITLE 选项设定了用浏览器打开该网页时的标题栏名称(后面的 TITLE 选项类似),CONTENTS 选项设定了目录文件名,FRAME 选项设定了框架文件名,PAGE 选项设定了页码文件名,PATH 选项设定了所有 HTML 文件的存储路径。也可以在指定各部分文件名时设定其存储路径,如“body = 'D:\class_body.html' (title = 'CLASS_BODY')”,对其他选项也可以做类似的修改。图 3-3 中网页内容分为三个部分:左上是目录文件内容,左下是页码文件内容,右面是主体文件内容,文件标题显示是 CLASS_FRAME,说明浏览器打开的是输出结果的框架文件。

• ODS PDF 语句

该语句控制 PDF 目标的打开、管理和关闭等操作,生成 PDF 文档来存储输出结果,可以使用 Adobe Acrobat 等软件来浏览这种格式的文档,属于第三方文件格式类。该语句使用格式如下:

```
ODS PDF < (<ID = >identifier) > <action>;  
ODS PDF < (<ID = >identifier) > <option(s)>;
```

如果不填写 ACTION 和 OPTIONS, ODS 会默认打开 PDF 目标并创建 PDF 输出。
ACTION 主要包括 CLOSE、EXCLUDE、SELECT、SHOW 这四个选项, 具体功能请参考本章相关内容。
ODS PDF 为用户提供了很多选项, 便于用户根据需求设计出想要的打印结果。表 3-2 总结了一些常用选项及其功能。

表 3-2 ODS PDF 常用选项及其功能

| 选 项 | 功 能 | 选 项 | 功 能 |
|----------------|--------------------------|-------------|-------------------|
| AUTHOR = | 指定输出文档的作者 | ID = | 同时运行同一输出目标的不同选项设定 |
| BACKGROUND = | 指定文本中是否打印背景色 | NEWFILE = | 在指定的起始点创建新文件 |
| BOOKMARKLIST = | 指定是否生成和显示 PDF 文档的书签 | STARTPAGE = | 控制分页 |
| COLOR = | 指定输出的颜色体制 | STYLE = | 指定用于输出的样式定义 |
| COLUMNS = | 指定每页输出的列数 | SUBJECT = | 指定输出文档的主题 |
| COMPRESS = | 控制 PDF 文件的压缩程度, 以降低其文件大小 | TEXT = | 指定文档中插入的文本 |
| CSSSTYLE = | 指定作用于输出的串接样式表 | TITLE = | 指定输出文档的标题 |
| DPI = | 指定输出文件中的图像分辨率 | UNIFORM = | 保证单个表格在页与页之间的一致性 |
| FILE = | 指定包含输出的文件 | | |

另外一种生成 PDF 输出的方式是利用 ODS PRINTER 语句, 通过该语句可以为各种打印格式目标进行输出, 还包括 PS 目标和 PCL 目标。
例如“ods printer ps style = d3d file = 'd3dstyle. ps';”语句生成了样式为 D3D、格式为 PS 的输出文档, 其他 PRINTER 目标格式文件也可以用类似语句生成。

【例 3-8】 利用 ODS PDF 语句生成完整的 PDF 文档。设程序名为 SASTJFX3_8. sas。

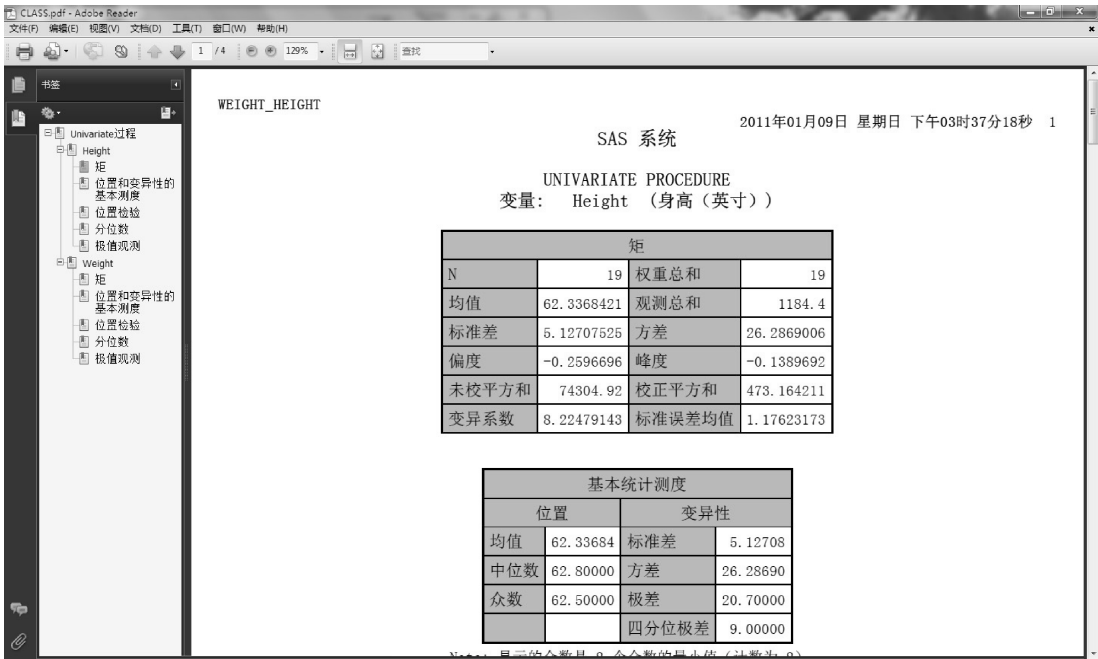
```
ods pdf  
file = 'D:\CLASS.pdf'  
text = 'WEIGHT_HEIGHT'  
startpage = yes  
title = 'CLASS_WEIGHT_HEIGHT'  
author = 'SABER'  
subject = 'CLASS'  
bookmarklist = show;  
proc univariate data = sashelp.class;  
var height weight;  
run;ods pdf close;quit;
```

运行结果如图 3-4 所示。
该例输出了一个完整形式的 PDF 文档, 文中设置了诸多选项, 主要完成了这样几个功能: “file = ”指定了文件名及其存储路径; “text = ”指定了首页左上角出现的文本内容; “startpage = ”控制页中断, “YES”在某些过程内部或者过程开始时另起一页, “NO”在某些过程内部或者过程开始时不翻页, 即便过程代码有翻页请求, 只有当一页填满时或者指定“startpage = now”才另起一页; “title = 、author = 、subject = ”分别对应了文档属性中的标题、作者和主题, 添加关键字可以用“keywords = ”选项实现; “bookmarklist = ”控制书签的生成与显示, 默认是“SHOW”, 例中代码可以省略此行。在运行结果的 PDF 截图中, 左侧是书签, 右侧是输出正

文，读者可以尝试改动和添加 ODS PDF 各种选项来观察输出文档格式的差异，从而得到令自己满意的打印文档。



(a)



(b)

图 3-4 UNIVARIATE 过程输出结果的开始部分之表现二

● ODS RTF 语句

该语句控制 RTF 目标的打开、管理和关闭等操作，生成 MS WORD 可识别的 RTF 文档，属于第三方文件格式类。该语句使用格式如下：

```
ODS RTF < (< ID = > identifier) > < action >;  
ODS RTF < (< ID = > identifier) > < option(s) >;
```

ACTION 主要包括 CLOSE、EXCLUDE、SELECT、SHOW 这四个选项，具体功能请参考本章相关内容。

ODS RTF 为用户提供了很多选项，便于用户根据需求设计出想要的打印结果。表 3-3 总结了一些常用选项及其功能。

用户可以使用各种 ODS RTF 选项来调整 RTF 目标，其中“FILE =”选项会自动关闭已经打开的 RTF 目标以及与其关联的所有文件，同时打开一个新的 RTF 目标，并创建新的 RTF 文件。

表 3-3 ODS RTF 常用选项及其功能

| 选 项 | 功 能 | 选 项 | 功 能 |
|-------------|-----------------------|---------------|--------------------|
| AUTHOR = | 指定输出文档的作者 | KEEPN NOKEEPN | 控制表格跨页时的处理 |
| BODYTITLE | 指定标题和脚注加入正文，而不放入页眉和页脚 | NEWFILE = | 在指定的起始点创建新文件 |
| COLUMNS = | 指定每页输出的列数 | PATH = | 指定外部文件的位置 |
| CONTENTS | 生成目录页 | SASDATE | 输出 SAS 时间和日期 |
| CSSSTYLE = | 指定作用于输出的串接样式表 | STARTPAGE = | 控制分页 |
| FILE = | 指定包含输出的文件 | STYLE = | 指定用于输出 RTF 文件的样式定义 |
| ID = | 同时运行同一输出目标的不同选项设定 | TEXT = | 指定文档中插入的文本 |
| IMAGE_DPI = | 指定图形输出分辨率 | TITLE = | 指定输出文档的标题 |

【例 3-9】 利用 ODS RTF 语句生成完整的 RTF 文档。设程序名为 SASTJFX3_9. sas。

```
ods rtf
file = 'CLASS.rtf' contents toc_data
text = 'WEIGHT_HEIGHT'
startpage = yes
title = 'CLASS_WEIGHT_HEIGHT'
author = 'SABER'
path = 'D:\'
bodytitle keepn sasdate;
proc univariate data = sashelp.class;
var height weight;
run;
ods rtf close;quit;
```

运行结果如图 3-5 所示。

该例输出了一个完整形式的 RTF 文档，文中设置了诸多选项，主要完成了这样几个功能：“file = ”指定了输出的 RTF 文件名；“contents”设定生成目录页；“toc_data”指示 ODS 把目录数据插入到 RTF 文档中；“text = ”指定了首页左上角出现的文本内容；“startpage = ”控制页中断，“YES”在某些过程内部或者过程开始是另起一页；“title = 、author = ”分别对应了文档属性中的标题和作者；“path = ”指定了输出文档的存储路径；“bodytitle”指定 SAS 标题和脚注写入正文，去掉此选项时，图 3-5(b)中的从“SAS 系统”到“(身高(英寸))”这三行字会进入页眉；“keepn”设定只有当整张表格无法在同一页显示完整时，才允许表格跨页；“sasdate”在文档中输出 SAS 时间和日期。图 3-5(a)给出的是目录页，默认目录内容是隐藏的，需要手动显示目录，MS Word 2007 的操作是“引用→更新目录”；图 3-5(b)给出的是正文的首页内容。读者可以尝试改动和添加 ODS RTF 各种选项来观察输出文档格式的差异，从而得到令自己满意的结果文档。

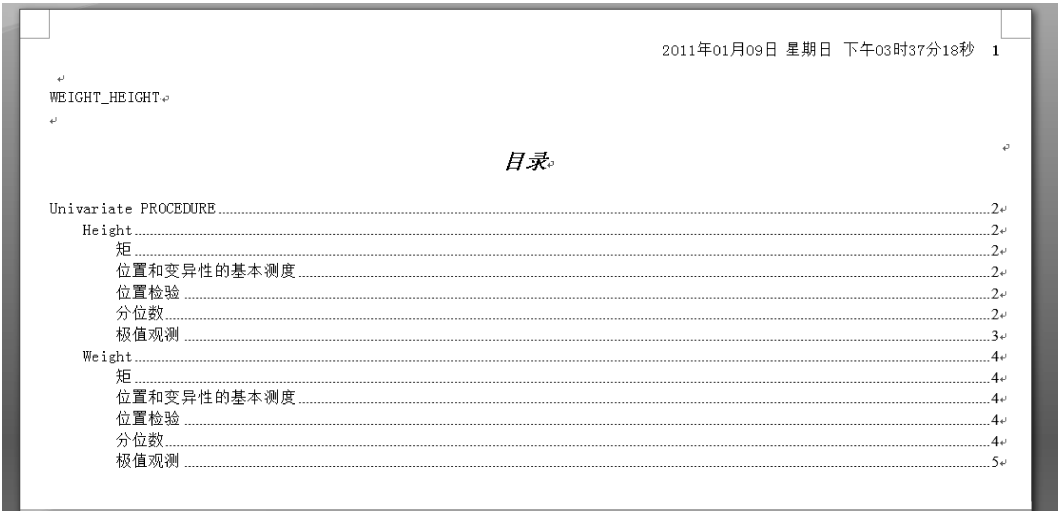
6. ODS OUTPUT 语句

该语句可以利用输出对象生成 SAS 数据，并管理 OUTPUT 目标的选择列表和剔除列表。该语句使用格式如下：

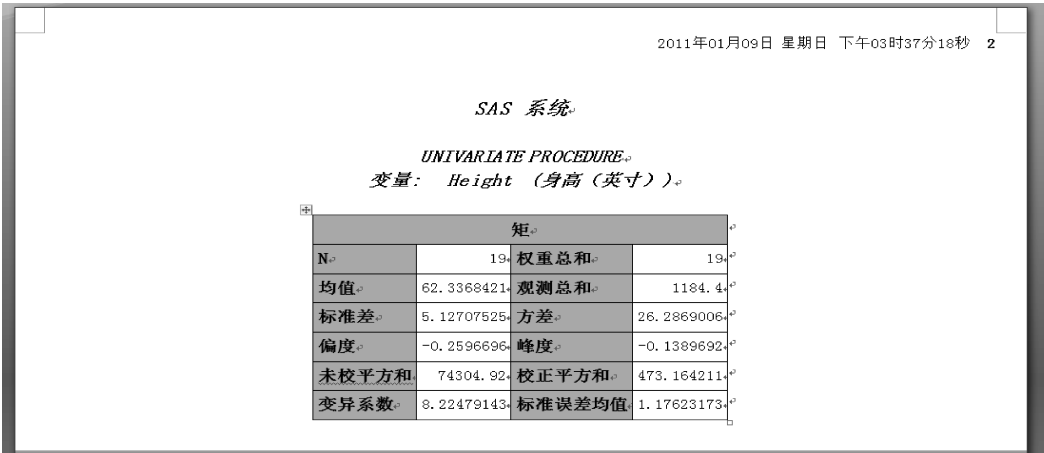
```
ODS OUTPUT action;
ODS OUTPUT data = set - definition(s);
```


ACTION 主要包括 CLEAR(剔除所有输出对象)、CLOSE(关闭 OUTPUT 目标)和 SHOW(在 SAS 日志中显示当前 OUTPUT 目标的选择列表和剔除列表)。

data-set-definition(s) 常见形式: output-object-specification < =data-set >



(a)



(b)

图 3-5 UNIVARIATE 过程输出结果的开始部分之表现三

这里不再强调输出对象的指定方法,前文已经做过详细的介绍,这里可以设定 MATCH_ALL 选项为每个输出对象都创建数据集。“data-set”位置填写指定的输出数据集名称,可以是一级名称或者二级名称。如果只创建一个输出数据集,那么 SAS 会直接使用指定的名称。如果使用了 MATCH_ALL 选项创建多个数据集,那么 SAS 会对创建的第一个数据集使用指定名称,随后的数据集名称由指定的名称加上数字编号组合而成(从第二个数据集开始名称依次为 data-set1, data-set2, data-set3, …)。

默认情况下,ODS OUTPUT 语句会把具有相同输出路径的输出对象合并到一个数据集中,不考虑任何输出对象之间变量的冲突。这里的“相同的输出路径”可以是两输出对象名称相同、部分路径相同、部分路径标签相同等多种情况。对于后一个数据集独有的变量,在前一个数据集的观测中,该变量值将被认为是缺失;类似地,对于前一个数据集独有的变量,在随后

的数据集观测中，其值也将被认为是缺失。如果两个数据集创建了一个同名不同类型的变量，那么后者的变量将被添加数字后缀来重命名。

【例 3-10】 利用 ODS OUTPUT 语句输出数据集。设程序名为 SASTJFX3_10.sas。

```
ods trace on/label listing;
ods output
Univariate.Height.Moments = Height_Moments_output (label = 'Height')
Univariate.Height.Moments = Height_Moments_partoutput (drop = cValue1 cValue2 label = 'Height_part')
Moments = Overall_Moments_output (keep = VarName Label1
nValue1 Label2 nValue2 label = 'Height_Weight');
proc univariate data = sashelp.class;
var height weight;
run;
ods output close;
ods trace off;quit;
```

运行结果如图 3-6 所示。

| VIEWTABLE: Height | | | | | | | |
|-------------------|---------|--------|------------|-----------|--------|------------|-------------|
| | VarName | Label1 | cValue1 | nValue1 | Label2 | cValue2 | nValue2 |
| 1 | Height | N | 19 | 19.000000 | 权重总和 | 19 | 19.000000 |
| 2 | Height | 均值 | 62.336842 | 62.336842 | 观测总和 | 1184.4 | 1184.400000 |
| 3 | Height | 标准差 | 5.12707525 | 5.127075 | 方差 | 26.2869006 | 26.286901 |
| 4 | Height | 偏度 | -0.2596696 | -0.259670 | 峰度 | -0.1389692 | -0.138969 |
| 5 | Height | 未校正平方和 | 74304.92 | 74305 | 校正平方和 | 473.164211 | 473.164211 |
| 6 | Height | 变异系数 | 8.22479143 | 8.224791 | 标准误差均值 | 1.17623173 | 1.176232 |

| VIEWTABLE: Height_part | | | | | |
|------------------------|---------|--------|-----------|--------|-------------|
| | VarName | Label1 | nValue1 | Label2 | nValue2 |
| 1 | Height | N | 19.000000 | 权重总和 | 19.000000 |
| 2 | Height | 均值 | 62.336842 | 观测总和 | 1184.400000 |
| 3 | Height | 标准差 | 5.127075 | 方差 | 26.286901 |
| 4 | Height | 偏度 | -0.259670 | 峰度 | -0.138969 |
| 5 | Height | 未校正平方和 | 74305 | 校正平方和 | 473.164211 |
| 6 | Height | 变异系数 | 8.224791 | 标准误差均值 | 1.176232 |

| VIEWTABLE: Height_Weight | | | | | |
|--------------------------|---------|--------|------------|--------|-------------|
| | VarName | Label1 | nValue1 | Label2 | nValue2 |
| 1 | Height | N | 19.000000 | 权重总和 | 19.000000 |
| 2 | Height | 均值 | 62.336842 | 观测总和 | 1184.400000 |
| 3 | Height | 标准差 | 5.127075 | 方差 | 26.286901 |
| 4 | Height | 偏度 | -0.259670 | 峰度 | -0.138969 |
| 5 | Height | 未校正平方和 | 74305 | 校正平方和 | 473.164211 |
| 6 | Height | 变异系数 | 8.224791 | 标准误差均值 | 1.176232 |
| 7 | Weight | N | 19.000000 | 权重总和 | 19.000000 |
| 8 | Weight | 均值 | 100.026315 | 观测总和 | 1900.500000 |
| 9 | Weight | 标准差 | 22.773933 | 方差 | 518.652047 |
| 10 | Weight | 偏度 | 0.183351 | 峰度 | 0.683365 |
| 11 | Weight | 未校正平方和 | 199436 | 校正平方和 | 9335.736842 |
| 12 | Weight | 变异系数 | 22.767942 | 标准误差均值 | 5.224699 |

图 3-6 UNIVARIATE 过程输出结果的开始部分之表现四

该例中用了两种方式引用输出对象，一共构建了三个数据集，前两个数据集用的都是输出对象的路径来做引用，第三个数据集用的是输出对象的名称来做引用。第一个数据集在指定了数据集名称之后还加了标签，但这个标签不会影响到跟踪记录里该输出对象的标签；第二个数据集额外用了“drop =”选项，表明结果数据集要去掉 DROP 指定的变量，类似的选项还有“where =”、“keep =”等。可以看到，对比第一个数据集，第二个数据集少了两个列变量；第三个数据集用的是输出对象的名称，并使用了“keep =”选项，而 HEIGHT 和 WEIGHT 变量都有一个名称为 MOMENTS 的输出对象，因此 ODS 会自动把这些同名的对象合并到一个数据集中去，从而得到第三个数据集的输出结果。我们要特别注意 ODS 会自动合并相同路径的数据集，防止发生意外的结果。

3.1.4 SAS ODS 的应用

本章意在强调 ODS 的应用，而不是指导读者怎么判别资料类型并选择分析方法。本章程

序力求简化输出结果，剔除掉在实际分析中无须关注的输出对象，只保留我们关注的分析结果，从而帮助读者加深对 ODS 应用的理解。本章的程序只是给出一种输出结果的建议模式，在分析问题时要根据实际情况灵活多变，搞清楚哪些输出结果是我们所关注的，也就是说要具体问题具体分析，从而达到既简化了输出结果，也没有丢失关键输出指标的目的。

1. 输出定量资料 t 检验结果

【例 3-11】 一个小麦新品种经过 6 代选育，从第 5 代(A 组)中抽出 10 株，株高为 66、65、66、68、62、65、63、66、68、62(cm)，又从第 6 代(B 组)中抽出 10 株，株高为 64、61、57、65、65、63、62、63、64、60(cm)。问：株高性状是否已经达到稳定？设程序名为 SASTJFX3_11.sas。

问题分析：该资料涉及两个组，每组随机抽取 10 个数据，测量指标为“株高”，属于成组设计一元定量资料。若该定量资料满足参数检验前提条件，可进行成组设计一元定量资料 t 检验，否则，应进行成组设计一元定量资料符号秩和检验。

程序实现：

```
DATA SASTJFX3_11;
  INPUT g $ n ;
  DO i = 1 to n;
    INPUT x @@ ; OUTPUT; END;
  CARDS;
  A 10
  66 65 66 68 62 65 63 66 68 62
  B 10
  64 61 57 65 65 63 62 63 64 60
  ;
RUN;

ods output
  TestsForNormality=TFNormality
  (where = (TestLab = 'W') drop =
  Test pSign);
PROC UNIVARIATE data = SASTJFX3_11
  NORMAL;
  VAR x;
  BY g;
RUN;
```

```
ods output
  Equality=EL
  TTests = TT (where = (Variances
    = 'Equal'));
PROC TTEST data = SASTJFX3_11
  COCHRAN;

  CLASS g;
  VAR x;
RUN;

ods html;
%macro printf(dataset);
proc print data = &dataset;
title "Dataset For &dataset";
run;
%mend;
%printf(TFNormality);
%printf(EL); %printf(TT);
ods _all_ close;
quit;
```

运行结果：

Dataset For TFNormality

| Obs | g | VarName | TestLab | Stat | pType | pValue |
|-----|---|---------|---------|----------|--------|--------|
| 1 | A | x | W | 0.902418 | Pr < W | 0.2329 |
| 2 | B | x | W | 0.899644 | Pr < W | 0.2171 |

Dataset For EL

| Obs | Variable | Method | NumDF | DenDF | FValue | ProbF |
|-----|----------|----------|-------|-------|--------|--------|
| 1 | x | Folded F | 9 | 9 | 1.31 | 0.6902 |

Dataset For TT

| Obs | Variable | Method | Variances | tValue | DF | Probt |
|-----|----------|--------|-----------|--------|----|--------|
| 1 | x | Pooled | Equal | 2.57 | 18 | 0.0193 |

这里着重解释 ODS 部分。第一个 ODS OUTPUT 语句是将正态性检验结果输出到数据集，对于“TestsForNormality”名称的输出对象有两个，分别对应 A 组与 B 组，所以 ODS 语句会自动将这两个同名的数据集合并到一起。另外，我们在这里使用了 WHERE 选项，目的是只保留 W 统计量，使用 DROP 选项去掉了两个列变量，使结果看起来更简洁；第二个 ODS OUTPUT 语句将方差齐性检验和 t 检验结果输出到数据集；最后利用宏将上面生成的三个数据集以网页格式输出，当然读者也可根据喜好输出成任意其他格式。利用 ODS OUTPUT 语句生成数据集的关键是弄清楚指定输出对象的名称、路径等信息，读者可以利用 ODS TRACE 语句来查看各输出对象的基本信息。

2. 输出定量资料非参数检验结果

【例 3-12】 现测得铅作业与非铅作业工人的血铅值(/100g)如表 3-4 所示，检验其差别。设程序名为 SASTJFX3_12. sas。

问题分析：该资料涉及两个组，测量指标为“血铅值”，属于成组设计一元定量资料。若该定量资料满足参数检验前提条件，可进行成组设计一元定量资料 t 检验，否则，应进行成组设计一元定量资料符号秩和检验。

程序实现：

表 3-4 两组工人的血铅值

| 非铅作业组 | 铅作业组 |
|--------|-------|
| 5 | 17 |
| 5 | 18 |
| 6 | 20 |
| 7 | 34 |
| 9 | 43 |
| 12 | 44 |
| 13 | 45 |
| 15 | . |
| 18 | . |
| 21 | . |
| N = 10 | N = 7 |

```
DATA SASTJFX3_12;
INPUT g $ n;
DO i =1 to n;
INPUT x @@ ;
OUTPUT;
END;
CARDS;
A 10
5 5 6 7 9 12 13 15 18 21
B 7
17 18 20 34 43 44 45
;
RUN;
ods output
TestsForNormality=TFNormality
(where = (TestLab = 'W') drop =Test pSign);
PROC UNIVARIATE data = SASTJFX3_12
NORMAL;
VAR x;BY g;
RUN;
ods output
Equality=EL;
PROC TTEST data = SASTJFX3_12 COCHRAN;
CLASS g; VAR x;
RUN;
ods output
WilcoxonTest =WT (where = (Name1 = 'Z_WIL' or Name1 = 'P2_WIL'));
PROC NPARIWAY data = SASTJFX3_12 WILCOXON;
```

```

CLASS g; VAR x;
RUN;
ods rtf startpage=no
bodytitle keepn;
%macro printf(dataset);
proc print data=&dataset noobs;
title "Dataset For &dataset";
run;
%mend;
%printf(TFNormality);%printf(EL);%printf(WT);
ods _all_ close;
quit;

```

运行结果:

Dataset For TFNormality

| g | VarName | TestLab | Stat | pType | pValue |
|---|---------|---------|----------|--------|--------|
| A | x | W | 0.919515 | Pr < W | 0.3529 |
| B | x | W | 0.814931 | Pr < W | 0.0574 |

Dataset For EL

| Variable | Method | NumDF | DenDF | FValue | ProbF |
|----------|----------|-------|-------|--------|--------|
| x | Folded F | 6 | 9 | 5.24 | 0.0278 |

Dataset For WT

| Variable | Name1 | Label1 | cValue1 | nValue1 |
|----------|--------|---------------------|---------|----------|
| x | Z_WIL | Z | 2.9313 | 2.931295 |
| x | P2_WIL | Two - Sided Pr > Z | 0.0034 | 0.003376 |

第一个 ODS OUTPUT 语句是将正态性检验结果输出到数据集,与上一个例子的选项设定相同;第二个 ODS OUTPUT 语句是将方差齐性检验结果输出到数据集,可以看到,该资料两组数据不满足方差齐性,因此我们需要关注非参数检验结果;第三个 ODS OUTPUT 语句将“WilcoxonTest”输出到数据集,同时用 WHERE 语句只保留了 Z 统计量 and 对应双侧检验 P 值;最后利用宏将上面生成的三个数据集以 WORD 格式输出,其中 ODS RTF 语句的相关选项在之前的内容已经做过详细介绍,读者可自行回顾。

3. 输出定量资料方差分析结果

【例 3-13】 从津丰小麦 4 个品系中分别随机抽取 10 株,测量其株高(cm),数据如下所示,问:不同品系津丰小麦的平均株高之间的差别是否具有统计学意义?设程序名为 SAS-TJFX3_13.sas。

品系 0-3-1: 63、65、64、65、61、68、65、65、63、64

品系 0-3-2: 56、54、58、57、57、57、60、59、63、62

品系 0-3-3: 61、61、67、62、62、60、67、66、63、65

品系 0-3-4: 53、58、60、56、55、60、59、61、60、59

问题分析：该资料涉及四个组，只有一个小麦品系因素，测量指标为“株高”，属于单因素四水平设计一元定量资料。若该定量资料满足参数检验前提条件，可进行单因素多水平设计定量资料一元方差分析；否则，应进行单因素多水平设计一元定量资料 Kruskal-Wallis 秩和检验。

程序实现：

```
DATA SASTJFX3_13;
INPUT GROUP $ N; DO i =1 TO N;
INPUT x @@ ; OUTPUT; END;
CARDS;
GROUP1 10
63 65 64 65 61 68 65 65 63 64
GROUP2 10
56 54 58 57 57 57 60 59 63 62
GROUP3 10
61 61 67 62 62 60 67 66 63 65
GROUP4 10
53 58 60 56 55 60 59 61 60 59
;
RUN;
ods output
TestsForNormality = TFNormality
(where = (TestLab = 'W') drop = Test
pSign);
PROC UNIVARIATE data = SASTJFX3_13;
NORMAL;
VAR x;
BY GROUP;
RUN;
```

```
ods output
HOVFTest = HT (where = (Source =
'GROUP'))
ModelANOVA = MANOVA
MCLines = SNK (drop = EqGROUP1 -
EqGROUP4);
PROC glm data = SASTJFX3_13;
CLASS GROUP;
MODEL X = GROUP/SS3;
MEANS GROUP / HOVTEST SNK ;
RUN;
ods pdf startpage = yes;
%macro printf(dataset);
proc print data = &dataset noobs;
title "Dataset For &dataset";
run;
%mend;
%printf(TFNormality);
%printf(HT);
%printf(MANOVA);%printf(SNK);
ods _all_ close;
ods listing;
quit;
```

运行结果：

Dataset For TFNormality

| GROUP | VarName | TestLab | Stat | pType | pValue |
|--------|---------|---------|----------|--------|--------|
| GROUP1 | x | W | 0.92112 | Pr < W | 0.3664 |
| GROUP2 | x | W | 0.955358 | Pr < W | 0.7319 |
| GROUP3 | x | W | 0.894252 | Pr < W | 0.1892 |
| GROUP4 | x | W | 0.884065 | Pr < W | 0.1452 |

Dataset For HT

| Effect | Dependent | Method | Source | DF | SS | MS | FValue | ProbF |
|--------|-----------|--------|--------|----|---------|---------|--------|--------|
| GROUP | x | LV | GROUP | 3 | 88.0627 | 29.3542 | 0.68 | 0.5705 |

Dataset For MANOVA

| Dependent | HypothesisType | Source | DF | SS | MS | FValue | ProbF |
|-----------|----------------|--------|----|------------|-------------|--------|--------|
| x | 3 | GROUP | 3 | 323.475000 | 107.8250000 | 17.52 | <.0001 |

Dataset For SNK

| Effect | Dependent | Method | Line1 | Mean | N | Level |
|--------|-----------|--------|-------|--------|----|--------|
| GROUP | x | SNK | A | 64.300 | 10 | GROUP1 |
| GROUP | x | SNK | A | — | — | |
| GROUP | x | SNK | A | 63.400 | 10 | GROUP3 |
| GROUP | x | SNK | | — | — | |
| GROUP | x | SNK | B | 58.300 | 10 | GROUP2 |
| GROUP | x | SNK | B | — | — | |
| GROUP | x | SNK | B | 58.100 | 10 | GROUP4 |

第一个 ODS OUTPUT 语句是将正态性检验结果输出到数据集，与上一个例子的选项设定相同，可以看到，四组数据的正态性检验结果被合并到一个数据集中，省去了我们分别查看各组结果的麻烦；第二个 ODS OUTPUT 语句是将方差齐性检验、方差分析以及两两比较的结果输出到数据集，可以看到，该资料四组数据是满足参数检验前提条件的，因此我们需要关注参数检验的结果；最后利用宏将上面生成的几个数据集以 PDF 格式输出，便于读者打印相关结果。

4. 输出定性资料卡方检验结果

【例 3-14】 65 例胰腺囊肿患者，按治疗方法分为 4 组，经皮置管引流 29 例，外引流 14 例，内引流 12 例，胰腺切除 10 例。观察结果，数据见表 3-5，考察各组间频数分布是否有差异。设程序名为 SASTJFX3_14.sas。

问题分析：该资料结果变量是治疗效果，是二值的，原因变量是治疗方法，是多值名义的，可按双向无序的 $R \times C$ 列联表资料进行分析。目的是比较原因变量各水平的频数分布情况，可以用一般卡方检验或 Fisher 精确检验来处理。

程序实现：

表 3-5 治疗方法和结果

| 治疗方法 | 例 数 | | |
|--------|-----|----|-----|
| | 结果： | 治愈 | 未治愈 |
| 经皮置管引流 | | 25 | 4 |
| 外引流 | | 11 | 3 |
| 内引流 | | 9 | 3 |
| 胰腺切除 | | 7 | 3 |
| 合计 | | 52 | 13 |

```
DATA SASTJFX3_14;
do a =1 to 4;
do b =1 to 2;
input f @@ ; output;
end;
end;
cards;
25 4
11 3
9 3
7 3
;
```

```
run;
ods html;
ods html select ChiSq FishersExact;
proc freq data = SASTJFX3_14;
weight f;
tables a* b/chisq;
exact fisher;
run;
ods html close;
quit;
```

运行结果：

FREQ PROCEDURE
“a * b”表的统计量

| 统计量 | 自由度 | 值 | 概率 |
|---|-----|--------|--------|
| 卡方 | 3 | 1.5286 | 0.6757 |
| 似然比卡方 | 3 | 1.5217 | 0.6773 |
| Mantel - Haenszel 卡方 | 1 | 1.4734 | 0.2248 |
| Phi 系数 | | 0.1534 | |
| 列联系数 | | 0.1516 | |
| Cramer V 统计量 | | 0.1534 | |
| WARNING: 38% 的单元格的期望计数比 5 小。 卡方可能不是有效检验。 | | | |

Fisher 精确检验

| | |
|---------|--------|
| 表概率(P) | 0.0139 |
| Pr <= P | 0.6161 |

本例采用 HTML 目标的输出格式，通过 SELECT 语句选择了两个关键的输出对象。结果中 WARNING 表明不满足卡方检验的前提条件，需要查看 Fisher 精确检验结果。

5. 输出定性资料秩和检验结果

【例 3-15】 某研究者为了比较三种糊剂治疗乳牙慢性根尖炎的治疗效果，将病情近似的患者 180 例，随机分为三组，碘仿糊剂组 61 例，Vitapax 糊剂组 56 例，自制糊剂组 63 例。治疗乳牙慢性根尖炎的效果见表 3-6。试对三组疗效的优劣进行比较。设程序名为 SASTJFX3_15.sas。

表 3-6 三种糊剂治疗乳牙慢性根尖炎的治疗效果比较

| 分 组 | 例 数 | | | |
|------------------|-------|-----|----|----|
| | 治疗效果: | 治愈 | 好转 | 未愈 |
| 碘仿糊剂组(A 组) | | 42 | 10 | 9 |
| Vitapax 糊剂组(B 组) | | 49 | 5 | 2 |
| 自制糊剂组(C 组) | | 54 | 5 | 4 |
| 合计 | | 145 | 20 | 15 |

问题分析：该资料结果变量是治疗效果，是多值有序变量，原因变量是分组情况，是多值名义变量，可按结果变量为有序变量的单向有序 R×C 列联表资料进行分析。目的是比较三组方法治疗效果的优劣，可以采用秩和检验来处理。

程序实现：

```
DATA SASTJFX3_15;
DO A=1 TO 3;
DO B=1 TO 3;
INPUT F @@ ; OUTPUT;
END;
END;
CARDS;
42 10 9
49 5 2
54 5 4
;

RUN;
ods html;
PROC NPAR1WAY data = SASTJFX3_15
WILCOXON;
CLASS A;
VAR B;
FREQ F;
RUN;
ods html close;
quit;
```


运行结果：

| The NPAR1WAY Procedure | | | | | |
|--|----|---------------|-------------------|------------------|------------|
| Wilcoxon Scores (Rank Sums) for Variable B | | | | | |
| Classified by Variable A | | | | | |
| A | N | Sum of Scores | Expected Under H0 | Std Dev Under H0 | Mean Score |
| 1 | 61 | 6178.00 | 5520.50 | 228.129590 | 101.278689 |
| 2 | 56 | 4700.50 | 5068.00 | 223.124920 | 83.937500 |
| 3 | 63 | 5411.50 | 5701.50 | 229.882773 | 85.896825 |
| Average scores were used for ties. | | | | | |
| Kruskal-Wallis Test | | | | | |
| Chi-Square | | 8.3949 | | | |
| DF | | 2 | | | |
| Pr > Chi-Square | | 0.0150 | | | |

本例采用 HTML 目标的输出格式，且没有对 NPAR1WAY 过程的输出对象进行筛选，读者可先根据 P 值得出三组疗效有差异的结论，然后再根据平均秩和对三组疗效作简单排序。

6. 输出定性资料相关分析结果

【例 3-16】 在固定矫治中龋齿活跃性变化的观察中，观察固定矫治中龋病活跃性变化。检测进行方丝弓固定矫正器治疗的 30 例不同年龄患者龋病活跃性度数，见表 3-7，考察年龄与 CAT 分度是否有相关性。设程序名为 SASTJFX3_16.sas。

问题分析：该资料结果变量是 CAT 分度，是多值有序变量，原因变量是年龄，是多值有序变量，可按双向有序且属性不同 $R \times C$ 列联表资料进行分析。目的是考察年龄与 CAT 分度是否有相关性，可以采用 Spearman 秩相关分析来处理。

程序实现：

| | |
|--|--|
| <pre>DATA SASTJFX3_16; DO A=1 TO 5; DO B=1 TO 4; INPUT F @@ ; OUTPUT; END; END; CARDS; 10 11 6 3 6 12 8 4 3 8 10 9 0 10 11 9 0 9 11 10 ;</pre> | <pre>RUN; ods html; ods html exclude VarInformation SimpleStats; PROC CORR SPEARMAN; VAR A B; FREQ F; RUN; ods html close; quit;</pre> |
|--|--|

运行结果：

| CORR PROCEDURE | | |
|--|----------------------------------|-----------------------|
| Spearman 相关系数, N = 150 当 H0: Rho = 0 时, Prob > r | | |
| | A | B |
| A | 1.00000 0.36361 <.0001 | 0.36361 <.0001 |
| B | 0.36361 <.0001 | 1.0000 |

本例采用 HTML 目标的输出格式, 通过 EXCLUDE 语句剔除了两个次要的输出对象。通过上表可以得到两变量有相关性的结论, 但相关系数取值较小, 没有实际应用价值。

7. 输出多重线性回归分析结果

【例 3-17】 某种水泥在凝固时放出的水化热 Y(卡/克)与水泥中的四种化学成分有关, 分别是 X1: 3CaO · Al₂O₃ 的含量(%), X2: 3CaO · AiO₂ 的含量(%), X3: 4CaO · Al₂O₃ · Fe₂O₃ 的含量(%), X4: 2CaO · SiO₂ 的含量(%). 共化验了 13 组数据(见表 3-8), 试分析这四种成分对水泥产生热量的影响。设程序名为 SASTJFX3_17.sas。

表 3-8 水泥在凝固时放出的热量与其四种成分的记录

| 编号 | X1 | X2 | X3 | X4 | Y | 编号 | X1 | X2 | X3 | X4 | Y |
|----|----|----|----|----|-------|----|----|----|----|----|-------|
| 1 | 7 | 26 | 6 | 60 | 78.5 | 8 | 1 | 31 | 22 | 44 | 72.5 |
| 2 | 1 | 29 | 15 | 52 | 74.3 | 9 | 2 | 54 | 18 | 22 | 93.1 |
| 3 | 11 | 56 | 8 | 20 | 104.3 | 10 | 21 | 47 | 4 | 26 | 115.9 |
| 4 | 11 | 31 | 8 | 47 | 87.6 | 11 | 1 | 40 | 23 | 34 | 83.8 |
| 5 | 7 | 52 | 6 | 33 | 95.9 | 12 | 11 | 66 | 9 | 12 | 113.3 |
| 6 | 11 | 55 | 9 | 22 | 109.2 | 13 | 10 | 68 | 8 | 12 | 109.4 |
| 7 | 3 | 71 | 17 | 6 | 102.7 | | | | | | |

问题分析: 该资料结果变量是水化热 Y, 是定量指标, 分析目的是四种成分对水泥产生热量的影响, 属于寻找因变量是如何随自变量变化而变化的数量依存关系, 因此可以采用多重线性回归分析分析来处理。

程序实现：

```
DATA SASTJFX3_17;
input id X1 -X4 Y @@ ;
cards;
1 7 26 6 60
2 1 29 15 52
3 11 56 8 20
4 11 31 8 47
5 7 52 6 33
6 11 55 9 22
7 3 71 17 6
8 1 31 22 44
9 2 54 18 22
10 21 47 4 26
11 1 40 23 34
12 11 66 9 12
13 10 68 8 12
```

```

;
run;
ods output SelectionSummary=summary(drop=PartialRSquare ModelRSquare Cp);
title 'Dataset of SelectionSummary';
proc reg data=SASTJFX3_17;
model Y=X1-X4/selection=stepwise;
model Y=X1-X4/selection=forward;
model Y=X1-X4/selection=backward;
run;
ods html;
proc print data=summary noobs;
run;
title;
ods html exclude NObs CollinDiag;
proc reg data=SASTJFX3_17;
model Y=X1 X2/collin collino int stb;
run;
title;
ods _all_ close;
ods listing;
quit;

```

运行结果:

Dataset of SelectionSummary

| Model | Dependent | Step | VarEntered | VarRemoved | NumberIn | FValue | ProbF |
|--------|-----------|------|------------|------------|----------|--------|--------|
| MODEL1 | Y | 1 | X4 | | 1 | 22.80 | 0.0006 |
| MODEL1 | Y | 2 | X1 | | 2 | 108.22 | <.0001 |
| MODEL1 | Y | 3 | X2 | | 3 | 5.03 | 0.0517 |
| MODEL1 | Y | 4 | | X4 | 2 | 1.86 | 0.2054 |
| MODEL2 | Y | 1 | X4 | | 1 | 22.80 | 0.0006 |
| MODEL2 | Y | 2 | X1 | | 2 | 108.22 | <.0001 |
| MODEL2 | Y | 3 | X2 | | 3 | 5.03 | 0.0517 |
| MODEL3 | Y | 1 | | X3 | 3 | 0.02 | 0.8959 |
| MODEL3 | Y | 2 | | X4 | 2 | 1.86 | 0.2054 |

The REG Procedure

Model: MODEL1

Dependent Variable: Y

| Analysis of Variance | | | | | |
|----------------------|----|----------------|-------------|---------|--------|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 2 | 2657.85859 | 1328.92930 | 229.50 | <.0001 |
| Error | 10 | 57.90448 | 5.79045 | | |
| Corrected Total | 12 | 2715.76308 | | | |

| | | | |
|----------------|----------|------------|--------|
| Root MSE | 2.40634 | R - Square | 0.9787 |
| Dependent Mean | 95.42308 | Adj R-Sq | 0.9744 |
| Coeff Var | 2.52175 | | |

| Parameter Estimates | | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|-----------------------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t | Standardized Estimate |
| Intercept | 1 | 52.57735 | 2.28617 | 23.00 | <.0001 | 0 |
| X1 | 1 | 1.46831 | 0.12130 | 12.10 | <.0001 | 0.57414 |
| X2 | 1 | 0.66225 | 0.04585 | 14.44 | <.0001 | 0.68502 |

| Collinearity Diagnostics(intercept adjusted) | | | | |
|---|------------|-----------------|-------------------------|---------|
| Number | Eigenvalue | Condition Index | Proportion of Variation | |
| | | | X1 | X2 |
| 1 | 1.22858 | 1.00000 | 0.38571 | 0.38571 |
| 2 | 0.77142 | 1.26199 | 0.61429 | 0.61429 |

这里着重解释 ODS 部分。第一个 ODS OUTPUT 语句是将变量筛选结果输出到数据集，例子中使用了三种变量筛选方法，会得到三个名为“SelectionSummary”的输出对象，所以 ODS 语句会自动将这三个同名的数据集合并到一起，并使用 DROP 选项去掉了三个列变量，由此我们可以很直观地发现三种方法筛选结果的不同；第二个 ODS HTML 语句将 OUTPUT 生成的数据集以及第二个 REG 过程产生的输出对象以网页格式输出，这里剔除掉了一些不必要的输出对象。读者由剩下的结果可以很快地判断模型是否有意义，做共线性诊断，判断变量作用强弱以及得出线性回归方程。

3.2 SAS 宏介绍

3.2.1 概述

本章介绍的宏工具是一种可以用来扩展 SAS 功能、减少普通输入文本工作量的 SAS 工具。可以使用宏工具给一段 SAS 程序或文本命名，并通过引用这个名称来使用这段程序或文本。SAS 系统的 MACRO 处理器可以让程序更简洁明了、更容易维护，帮助用户在使用 SAS 系统时更方便、更自动化。具体来说，它具有以下功能。

(1) 利用宏工具可以将一个变量、一段程序或者一个文本命名，供以后调用，是用于扩充和制作用户化 SAS 系统的工具。利用宏功能可以减少用户在完成一些重复任务时必须输入的文本量，从而使得程序简洁明了，便于调试与修改。当用户在某个 SAS 程序中使用宏语句时，这个宏语句便开始实现其所替代内容的功能。

(2) 获取 SAS 的系统信息。SAS 在启动时就创建了一些自动宏变量，用以存储当前 SAS 进程启动的日期、时间、版本号及其他信息，用户可以在任何情况下使用这些宏变量。

(3) 有条件地执行数据步和过程步。例如，每天提交一份病人情况的详细报告，每周一增加一份上周的汇总报告，使用宏功能每天运行同一个程序就可以实现上述任务。

(4) 开发交互式系统。使用 SAS 宏语言的 % WINDOW 语句及一些基本的编程语句可开发交互式用户界面。

(5) 产生与数据无关的 SAS 程序，但可展示与数据相关的结果。宏功能可保持 SAS 程序的独立性和移植性，定义的宏变量或宏语句可游离于数据步和过程步之外，一段宏程序在多种情况下均可运行，并得到期望的结果。

(6)在不同的 SAS 数据步和过程步之间传递数据。SAS 宏工具可以定义全局变量，该变量可在 SAS 的任何地方被引用，所以成为不同过程间传递数据最方便的手段。

(7)重复执行 SAS 程序代码。引用 SAS 宏变量或宏语句时，以符号%或&作为开始。在编译阶段，系统对程序逐词扫描，当遇到%或&开始的词，就启动宏语言处理器对此进行处理。

3.2.2 宏变量

在 SAS 程序中，宏变量是一种通过象征符号来动态更改文本的工具。用户可以事先将宏变量指定为某些文本，然后通过引用该变量来达到使用文本的目的。

宏变量值的最大长度限定为 32K 字符，它的长度通常是由赋值给它的文本决定的，由于文本内容的多变性，宏变量的长度也会随之变化，因此，在定义宏变量时一般不对其长度加以限定。宏变量只包含字符型数据，不过，当这个变量所包含的数据可以解释为数值时，宏功能中允许把这个变量作为数值进行计算。在程序编译过程中，宏变量的值通常保持不变，除非对其值进行明确的改变，对 SAS 数据集变量的操作不会影响到宏变量。

宏变量分为用户自定义宏变量和自动宏变量。用户可以在程序的任何地方定义和使用宏变量。

当一个宏变量被定义时，宏处理器自动将其添加到程序中的宏变量符号表中。如果宏变量是在宏定义以外创建的或是宏处理器自动产生的(SYSPBUFF 除外)的，该变量被放在每次 SAS 系统启动时都会产生的全局符号表中。如果宏变量是在宏程序内部定义的，而且没有声明是全局变量，则该变量被放在宏自身的符号表中，每次宏运行时都会更新这张表。

存放于全局符号表中的变量称为全局变量，其值可以被 SAS 程序的任何一部分调用(CARDS 以及 DATALINES 语句除外)，该变量在 SAS 整个运行期间均作用有效。SAS 程序的其他部分也可能产生全局宏变量，但是只有宏处理器本身产生的宏变量才被认为是自动宏变量。

存放于局部符号表中的变量称为局部宏变量，其值只有在执行定义它的宏时才可被引用，在这个宏以外的程序中该变量并不存在。

用户可以使用%PUT 语句查看当前 SAS 系统运行过程中所有可以利用的宏变量。

1. 宏变量的定义

用户可以根据需要自主创建宏变量，改变宏变量的值，以及定义它的作用范围。宏变量必须以字母或者下画线开头，后面跟字母或者数字。用户可以任意定义宏变量的名称，但不能与系统保留字冲突。建议不要使用 AF、DMS、SQL 和 SYS 等前缀，因为 SAS 软件常把它们用在自动宏变量中，用户要尽量避免自定义宏变量与自动宏变量之间的混淆。如果用户在定义时使用了不合法的变量名，在 SAS 的 log 中会提示错误信息。用时可参阅表 3-9。

表 3-9 宏功能中的保留字

| | | | | |
|----------|--------|--------|----------|----------|
| ABEND | END | LENGTH | QKUPCASE | SYSEVALF |
| ABORT | EVAL | LET | QSCAN | SYSEXEC |
| ACT | FILE | LIST | QSUBSTR | SYSFUNC |
| ACTIVATE | GLOBAL | LISTM | QSYSFUNC | SYSGET |
| BQUOTE | GO | LOCAL | QUOTE | SYSRPUT |
| BY | GOTO | MACRO | QUPCASE | THEN |
| CLEAR | IF | MEND | RESOLVE | TO |

| | | | | |
|----------|---------|----------|----------|---------|
| CLOSE | INC | PAUSE | RETURN | TSO |
| CMS | INCLUDE | NRSTR | RUN | UNQUOTE |
| COMANDR | INDEX | ON | SAVE | UNSTR |
| COPY | INFILE | OPEN | SCAN | UNTIL |
| DEACT | INPUT | PUT | STOP | UPCASE |
| DEL | KCMPRES | NRBQUOTE | STR | WHILE |
| DELETE | KINDEX | NRQUOTE | SYSCALL | WINDOW |
| DISPLAY | KLEFT | METASYM | SUBSTR | |
| DMIDSPLY | KLENGTH | QKCMRES | SUPERQ | |
| DMISPLIT | KSCAN | QKLEFT | SYMDEL | |
| DO | KSUBSTR | QKSCAN | SYMEXIST | |
| EDIT | KTRIM | QKSUBSTR | SYMGLOBL | |
| ELSE | KUPCASE | QKTRIM | SYMLOCAL | |

用户可以使用% PUT_ALL_查看所有自定义宏变量。

定义宏变量最简单的方式是使用宏程序语句% LET, 定义方式如下:

```
%let mac_var =Newdata;
```

mac_var 是宏变量名, Newdata 是该宏变量对应的值。宏变量值通常是一串字符, 可以包含任何字母、数字以及键盘上可找到的可打印符号, 字符间允许有空格。

如果定义的宏变量已经存在, 那么上面的操作将会导致原宏变量值的覆盖。

2. 宏变量的直接引用

宏变量被创建之后, 用户只要在变量名称前加一个与符号(&mac_var), 即可实现宏变量的引用。宏变量最大的方便之处就是, 用户可以在程序的任何地方加以引用。

【例 3-18】 引用宏变量 mac_var。设程序名为 SASTJFX3_18. sas。

```
%let mac_var =Newdata;
data;
%put &mac_var;
run;
```

上段程序等价于:

```
data;
put "Newdata";
run;
```

程序运行结果为将字符串 Newdata 打印出来。

为了能够在文本字符串中正确引用宏变量, 需要用双引号将文本括起来, 单引号中文本的宏变量不能被解析。

【例 3-19】 引用文本字符串中的宏变量。设程序名为 SASTJFX3_19. sas。

```
%let mac_var =SAS;
data;
put "data set of &mac_var";
put 'data set of &mac_var';
run;
```

第一个 put 的输出结果是 data set of SAS, 第二个 put 的输出结果是 data set of &mac_var。因此, 引用引号内的宏变量时必须用双引号将文本括起来。

用户可以根据需要多次引用宏变量，宏变量的值始终保持不变，直到用户对其进行更改。

【例 3-20】 多次引用宏变量。设程序名为 SASTJFX3_20.sas。

```
%let dsn = Newdata;
data temp;
    set &dsn;
run;
proc print;
    title "Subset of Data Set &dsn";
run;
```

每次 &dsn 出现时，SAS 程序自动将其替换为 Newdata，上段程序等价于：

```
DATA TEMP;
SET NEWDATA;
RUN;

PROC PRINT;
TITLE "Subset of Data SetNewData";
RUN;
```

如果引用了不存在的宏变量，SAS 的 log 中会提示错误信息。

很多情况下，用户可能面临着宏变量和前导文本或者末尾文本合并使用的问题（例如，DATA = PERSNL&YR. EMPLOYES,&YR 代替了两位数的年份值），以及两宏变量合并起来使用的情形（例如，&MONTH&YR）。用户可以用简洁易懂的宏变量名称代替该文本，通过使用这个共同的宏变量名称达到在多个地方使用这个文本的目的，而且，只要修改了宏定义，就间接地修改了所有该文本出现的位置。

【例 3-21】 宏变量和文本的合并使用。设程序名为 SASTJFX3_21.sas。

```
%let name = sales;
data new&name;
set save.&name;
run;
```

这段程序等价于：

```
DATA NEWSALES;
SET SAVE.SALES;
RUN;
```

某些情况下，用户需要引用宏变量作为前缀，但是仅仅将宏变量与文本连接起来写入程序，宏处理器是不能正确解析出宏变量的。假设用户仅定义了名为 name 的宏变量，而用户可能需要 &name1、&name2 如此引用，这样的写法会使宏处理器误认为要解析的是 name1、name2 两个宏变量，显然，系统会提示两个宏变量并不存在。在宏变量引用过程中，宏处理器逐字扫描文本，当遇到不能用于宏变量名中的字符时，该扫描动作停止，在此之前读入的文本信息会被解析为一个完整的宏变量，我们通过引入“.”作为划界符，来控制宏处理器的扫描范围。

【例 3-22】 宏变量与非宏变量文本的分隔。设程序名为 SASTJFX3_22.sas。

```
%let mac_var = name;
data;
    %put &mac_var.1;
```

```
%put &mac_var..1;
%put &mac_var^1;
run;
```

第一个 put 的输出结果为 name1；第二个 put 的输出结果为 name.1，这里解决了这样一个问题：如果“.”也是要输出文本中的一部分，我们只需要用两个划界符，前一个起到标志宏变量名结束的作用，后一个则是文本中的一部分；第三个 put 的输出结果为 name^1，因为“^”在宏变量名中不能使用，宏处理器不会认为“^”是宏变量名的一部分。

由第三个 put 的输出结果可知，下面两语句等价：

```
%put &mac_var^1;
%put &mac_var.^1;
```

3. 宏变量值的显示

最简单地显示宏变量值的方式是用 % PUT 语句，结果会输出在 SAS 的 log 窗口中。

【例 3-23】 显示宏变量的值。设程序名为 SASTJFX3_23.sas。

```
%let mac_var1=1;
%let mac_var2=2;
%let mac_var3=3;
data;
%put &mac_var1 + &mac_var2 = &mac_var3;
run;
```

输出结果为：1 + 2 = 3。

4. 宏变量值的改变

宏变量值的改变与宏变量的定义类似，只要再次使用 % LET 语句对宏变量进行赋值操作即可。

【例 3-24】 改变宏变量的值。设程序名为 SASTJFX3_24.sas。

```
%let mac_var=1;
data;
%put &mac_var;
%let mac_var=2;
%put &mac_var;
run;
```

上面的例子显示了宏变量的值由 1 变为 2 的操作流程。

5. 宏变量的间接引用

第 2 小节讲的都是对宏变量的直接引用方式，而在某些情况下，可能需要用户采取对宏变量的间接引用方式。假设我们定义了一系列宏变量 name1, name2, ..., name10，又定义了一个宏变量 n，用户可能需要根据 n 的变化来引用不同的宏变量 name，也就是说当 n 为 3 时，需要引用 name3，当 n 为 8 时，需要引用 name8。根据上面的前提条件，我们可能会想到用下面的语句来实现：

```
%put &name&n;
```

然而，SAS 将会提示错误信息，指出 name 不是一个宏变量，因为宏处理器会试图去解析 name 和 n，然后再将它们的值连接起来。我们做如下修改：


```
%put &&name&n;
```

这里，我们添加了一个“&”，让宏处理器多扫描一次以确认是宏变量的间接引用。上述语句解析的基本流程是，先将 && 解析为一个 &，name 作为文本，&n 被解析为数字（设为 3），这样该语句经过一次扫描就解析为 &name3，之后宏处理器开始解析宏变量 name3，返回 name3 所对应的值。

【例 3-25】 宏变量的间接引用。设程序名为 SASTJFX3_25.sas。

```
%let name1 = 1;
%let name2 = 2;
%let name3 = 3;
%macro test;
%do n = 1 % to 3;
%put &&name&n;
%end;
%mend test;
%test
```

上述程序运行结果就是 1 2 3。

宏处理器处理多个 & 的原则是，从左到右把每两个 & 解析成一个 &，然后解析后面的内容。

【例 3-26】 多个 & 时宏变量的引用。设程序名为 SASTJFX3_26.sas。

```
%let mac_var3 = result;
%let name = mac_var;
%let n = 3;
data;
%put &&&name&n;
run;
```

运行结果是在 SAS 的 log 窗口中输出 result。上述 put 语句解析的基本流程是，先将 && 解析为一个 &，&name 解析为 mac_var，&n 解析为 3，这样该语句经过一次扫描就解析为 &mac_var3，之后，宏处理器开始解析宏变量 mac_var3，返回 mac_var3 所对应的值 result。

6. 自动宏变量

当用户启动 SAS 系统时，宏处理器会自动创建一些宏变量用以保存 SAS 运行期间的相关信息。除了 SYSPBUFF 是局部变量外，其余自动宏变量都是全局变量。引用自动宏变量的方法与引用用户自定义宏变量的方法相同，即 & 后面紧跟自动宏变量名。

【例 3-27】 自动宏变量的引用。设程序名为 SASTJFX3_27.sas。

```
data;
%put "Today is &sysday, &sysdate9";
run;
```

sysday 返回的是 SAS 系统当前运行时间是星期几，sysdate9 返回的是 SAS 系统当前运行时间的日期，且年份用四位数表示。

自动宏变量分为两种状态，据此分为两类：一类是可读/写的自动宏变量，用户可以根据需要对这类宏变量赋值；另一类是只读的宏变量，用户不能对这一类宏变量做任何修改。

可以使用 % PUT _AUTOMATIC_ 查看所有可以利用的自动宏变量，见表 3-10。

表 3-10 按读/写状态划分的自动宏变量

| 状 态 | 变量名称 | 含 义 |
|------|-----------------|--|
| 可读/写 | SYSBUFFR | 接受随%INPUT 语句输入但宏处理器不能同该语句中任何一个变量匹配的文本 |
| | SYSCC | 操作环境代码 |
| | SYSCMD | 存放来自宏窗口命令行中最后一条不可识别的命令 |
| | SYSDEVIC | 当期画图设备的名称 |
| | SYSDMG | 反映对一个损坏的数据集采取措施的返回代码 |
| | SYSDSN | 存放最新创建的 SAS 数据集的完整名称, 库名和数据集名可分开显示, 形式为“库名 SAS 数据集名” |
| | SYSFILRC | FILENAME 语句设置的返回代码 |
| | SYSLAST | 存放最新创建的 SAS 数据集的完整名称, 形式为“库名.SAS 数据集名” |
| | SYSLCKRC | LOCK 语句设置的返回代码 |
| | SYSLIBRC | LIBNAME 语句设置的返回代码 |
| | SYSLOGAPPLNAME | 存放选项 LOGAPPLNAME 的值 |
| | SYSMMSG | 存放用户在宏窗口信息域中规定被显示的信息 |
| | SYS Parm | 存放系统选项 SYSPARM 规定的值 |
| | SYSBUFF | 宏参数值的文本 |
| | SYSRC | 各种系统有关的返回代码 |
| 只读 | SYSCHARWIDTH | 字符串宽度值 |
| | SYS DATE | SAS 系统当前运行的日期, 年份是两位数 |
| | SYS DATE9 | SAS 系统当前运行的日期, 年份是四位数 |
| | SYS DAY | SAS 系统当前运行时间是一周的哪一天 |
| | SYS ENCODING | SAS 系统当前编码方式名称 |
| | SYS ENV | SAS 系统运行在前景模式还是背景模式的指示符 |
| | SYS ERR | SAS 过程步和数据步设置的返回代码 |
| | SYS ERRORTEXT | 存放最近一次显示在 SAS log 中的错误信息 |
| | SYS HOSTNAME | 当前操作系统的用户名 |
| | SYS INDEX | 存放 SAS 系统当前运行期间已开始执行的宏个数 |
| | SYS INFO | 返回代码信息 |
| | SYS JOBID | SAS 当前工作或用户的 ID(随主机环境改变) |
| | SYS MACRONAME | 当前执行宏的名称 |
| | SYS MEN | 当前宏的执行环境 |
| | SYS NCPU | 当前 SAS 运算可能需要的处理器数量 |
| | SYS ODS PATH | ODS 中 PATH 变量的值 |
| | SYS PROCESSID | 当前 SAS 编译环境的 ID |
| | SYS PROCESSNAME | 当前 SAS 编译环境的名称 |
| | SYS PROCNAME | 当前编译的过程名 |
| | SYS SCP | 当前 PC 操作系统的缩写 |
| | SYS SCPL | 当前 PC 操作系统的名称 |
| | SYS SITE | 存放指定位置号码 |
| | SYS STARTID | 最近一条 STARTSAS 语句产生的 ID |
| | SYS STARTNAME | 最近一条 STARTSAS 语句产生的操作名称 |
| | SYS TIME | SAS 系统开始执行的时间 |
| | SYS USERID | 当前 SAS 操作的用户 ID 或注册号 |
| | SYS VER | 返回 SAS 软件的版本号 |
| | SYS VLONG | 返回 SAS 软件的版本号和维持水平的两位数年份 |
| | SYS VLONG4 | 返回 SAS 软件的版本号和维持水平的四位数年份 |
| | SYS WARNINGTEXT | 存放最近一次显示在 SAS log 中的警告信息 |

7. 全局宏变量

每一个宏变量都有自己的作用范围, 根据作用范围大小不同将宏变量分为全局宏变量和

局部宏变量。全局宏变量在整个 SAS 程序运行期间均有效，而且可以在程序的任何地方予以使用；局部宏变量只能在创建该变量的宏中使用。在这个宏之外，这个局部宏变量没有任何意义。

宏可以互相嵌套，对于定义在这些宏中的局部宏变量，我们要弄清楚它们的使用范围。例如，macro_var1 是定义在宏 macro1 中的宏变量，macro_var2 是定义在宏 macro2 中的宏变量，宏 macro2 是定义在宏 macro1 内部的，因此 macro_var1 是宏 macro1 和宏 macro2 的局部宏变量，而 macro_var2 只是 macro2 的局部宏变量，对于宏 macro1 没有任何意义。

本节着重介绍全局宏变量，下节着重介绍局部宏变量。

全局宏变量包括：

- 所有自动宏变量(SYSPBUFF 除外)；
- 在任何宏之外定义的宏变量；
- 用%GLOBAL 语句创建的宏变量；
- 由 CALL SYMPUT 程序创建的大部分宏变量。

【例 3-28】 简单创建全局宏变量的方法。设程序名为 SASTJFX3_28.sas。

```
%let mac_var =Newdata;
%macro mac;
%global mac_var1;
%let mac_var1 =Newdata1;
%mend mac;
%mac;
%put_global_;
run;
```

该例子用了两种创建全局宏变量的方法：一种是在任何宏之外创建宏变量(如变量 mac_var)；另一种是在宏之内创建宏变量，同时声明它是全局的作用范围，用户可以通过输出全局变量表来查看全局变量是否创建成功。

用户可以在 SAS 系统运行的任何时间创建全局宏变量，除了一些只读的自动宏变量外，这些全局宏变量可以随时根据用户需要进行修改。当一个全局宏变量的名称已经存在，而用户又在某个宏内定义了相同名称的宏变量(不是用%LOCAL 语句创建也不是宏参数)时，该变量并不是局部宏变量，该操作只相当于对这个已经存在的全局宏变量重新进行了赋值，用户在创建宏变量时要避免与已经存在的宏变量或者自动宏变量的名称相冲突。

引用全局宏变量时应注意以下问题：

- 当一个宏变量既存在于全局符号表内时，又存在于局部符号表内时，如果在定义它的宏内引用该变量，那么我们是将其作为局部宏变量来引用的，原因是宏处理器在遇到 & 优先选择在局部符号表中查询该变量。
- 若一个全局宏变量是在数据步用 SYMPUT 子程序创建的，那么只能在该数据步运行结束之后才可以引用该变量，原因是 SYMPUT 子程序是在数据步结尾才起作用并创建宏变量的。

8. 局部宏变量

定义在宏内的宏变量都是局部宏变量。每个宏被定义之后都会创建属于自己的局部符号表，属于该宏的局部宏变量都会被自动添加到这个表内。当这个宏执行结束之后，所有属于该宏的局部宏变量都将不存在。

局部宏变量包括：

- 由%LOCAL 语句创建的宏变量；
- 与宏对应的宏参数；
- 由宏语句定义的宏变量(该变量不是已经存在的全局变量，或者没有被声明为%GLOBAL)。

【例3-29】 简单创建局部宏变量的方法。设程序名为 SASTJFX3_29.sas。

```
%let mac_var = Newdata;
%macro mac(mac_var1);
%let mac_var2 = Newdata2;
%put **** inside macro**** ;
%put &mac_var &mac_var1 &mac_var2;
%mend mac;
%mac (Newdata1);
%put **** outside macro**** ;
%put &mac_var &mac_var1 &mac_var2;
run;
```

上例中给出了两种创建局部宏变量的方法：一种是在宏内直接定义宏变量，另一种是将变量作为宏参数。上例还给出了局部宏变量与全局宏变量作用范围的区别，几个%PUT 的输出结果为：

```
**** inside macro****
Newdata Newdata1 Newdata2
**** outside macro****
```

WARNING:没有解析符号引用 MAC_VAR1。

WARNING:没有解析符号引用 MAC_VAR2。

```
Newdata &mac_var1 &mac_var2
```

可以看到，全局宏变量 mac_var 无论是在宏内还是宏外均能被正确解析，而两个局部宏变量 mac_var1 和 mac_var2 只有在它们所属的宏执行时才能被正确解析。

宏处理器对宏变量的创建、赋值以及解析都遵循这样的流程：当宏处理器收到一个对宏变量的操作请求后，宏处理器先判断这一请求是来自宏外的程序还是宏内的程序。假设请求是来自宏外的，宏处理器会检查要操作的宏变量是否存在于全局符号表内，如果存在则进行相应的操作，如果不存在则创建或赋值操作均转化为创建宏变量的操作，而解析操作会提示警告信息；假设请求是来自宏内的，宏处理器先从最内部的局部符号表开始查询该变量是否存在，如果存在则进行相应操作，不存在则继续查询同层次的局部符号表，如果同层次的局部符号表已全部搜索过一遍，则开始查询较外层的局部符号表，在所有局部符号表均搜索不到该变量的情况下，宏处理器最后会搜索全局符号表，以完成对宏变量的操作请求。

3.2.3 宏与宏参数

宏是一段由用户事先编辑好的程序，用户可以根据需要对宏进行调用。与宏变量类似，一般用宏来产生文本，但宏有额外的以下优势：

- 通过宏语句可以控制何时以及如何产生文本；
- 可以定义宏参数，方便用户重复使用宏，并且达到不同的输出目的。

1. 创建名为 mac 的宏

宏定义的基本格式为：

```
%MACRO 宏名;  
宏程序内容;  
%MEND 宏名;
```

定义一个宏时，以% MACRO 开始，以% MEND 结束，宏名要避免与其他的变量名或者关键字冲突。宏程序是由宏语句、宏引用、文本表达式以及恒定文本组合而成的。宏在替代复杂文本时，往往会显示出很大的优势。

【例 3-30】 创建名为 mac 的宏。设程序名为 SASTJFX3_30. sas。

```
%macro mac;  
%let mac_var = Newdata;  
%put &mac_var;  
%mend mac;  
run;
```

上面的例子创建了一个名为 mac 的宏，该宏创建了一个局部宏变量 mac_var，并输出了该宏变量对应的值。

2. 创建形如 mac(variable1, variable2, ...) 的宏

在很多情况，我们会觉得频繁使用% LET 语句定义宏变量显得很烦琐，不仅不利于程序简洁，而且也对各宏变量的意义区分增加了困难。这时，就需要用宏参数把宏变量和宏有机地结合起来。

我们把在% MACRO 语句中定义的并用圆括号括起来的宏变量称为宏参数。宏参数有这样几个优势：首先，可以减少我们使用% LET 语句的次数，其次，宏参数可以确保宏变量不会干扰到宏外的 SAS 程序，定义宏参数事实上就是创建局部宏变量，只有在定义它的宏执行时，该变量才存在。

【例 3-31】 创建形如 mac(variable1, variable2, ...) 的宏，设程序名为 SASTJFX3_31. sas。

```
%macro mac(var1,var2,var3);  
%put &var1 &var2 &var3;  
%mend mac;  
run;
```

上面的例子创建了一个具有三个参数的宏，在宏内部对三个参数的值进行了输出，下面将介绍给宏参数赋值的方法。

3. 给宏参数赋值

之所以选择定义宏参数，多数情况是为了在赋值和修改宏变量上能够大大简化程序，提高对宏的利用效率。下面介绍两种常用的给宏参数赋值的方法。

【例 3-32】 创建宏时直接赋值。设程序名为 SASTJFX3_32. sas。

```
%macro mac(var1 = data1, var2 = data2, var3 = data3);  
%put &var1 &var2 &var3;  
%mend mac;  
%mac;  
run;
```

上面的例子在创建宏的同时，给宏参数进行了赋值。

【例 3-33】 引用宏时进行赋值。设程序名为 SASTJFX3_33.sas。

```
%macro mac(var1,var2,var3);  
%put &var1 &var2 &var3;  
%mend mac;  
%mac(data1,data2,data3);  
run;
```

从这两个例子可以看出，两种不同的宏参数的赋值方法在程序编写上的区别。

3.2.4 宏的引用

上一节对宏和宏参数的定义做了简要的介绍，本节将集中介绍宏的引用以及宏函数的引用。

1. 引用名为 mac 的宏

引用宏的基本方法是在宏的名称前加一个百分号(%)。

【例 3-34】 引用名为 mac 的宏。设程序名为 SASTJFX3_34.sas。

```
%macro mac;  
%let var1=data1;  
%let var2=data2;  
%let var3=data3;  
%put &var1 &var2 &var3;  
%mend mac;  
%mac;  
run;
```

例中先定义了一个名为 mac 的宏，宏内创建了三个局部宏变量，我们用 %mac 来引用该宏，该操作的结果是将三个宏变量值输出到 SAS log 中。

【例 3-35】 多次引用相同的宏。设程序名为 SASTJFX3_35.sas。

```
%let var1=data_old1;  
%let var2=data_old2;  
%let var3=data_old3;  
%macro mac;  
%put &var1 &var2 &var3;  
%mend mac;  
%mac;  
%let var1=data_new1;  
%let var2=data_new2;  
%let var3=data_new3;  
%mac;  
run;
```

例中两次引用了名为 mac 的宏，但两次引用的结果并不相同。例中定义了三个全局宏变量，第一次引用之后对三个全局宏变量的值进行了修改，因此第二次引用该宏输出的是修改之后新的宏变量值。

如果我们定义了三个局部宏变量，并且在某个环节上要对这三个宏变量的值进行修改，若不引入宏参数，我们能想到的修改方式就是将宏重新定义一次，定义过程中对变量赋予新值，

若用户需要修改多次,那么程序将会变得无比烦琐,下面介绍的带参数的宏很好地解决了这一问题。

2. 引用形如 `mac(variable1, variable2, ...)` 的宏

宏参数并没有严格的个数限制,用户在定义宏时可以根据需要将频繁使用的变量设为宏参数,以便于随时修改变量的值。

【例 3-36】 定义并引用形如 `mac(variable1, variable2, ...)` 的宏,设程序名为 `SASTJFX3_36.sas`。

```
%macro mac(var1,var2,var3);
%put &var1 &var2 &var3;
%mend mac;
%mac(data_old1,data_old2,data_old3);
%mac(data_new1,data_new2,data_new3);
run;
```

我们利用带参数的宏可以很方便地为局部宏变量赋值,即便需要重复引用,用户也可以在引用的同时完成赋值。

3. 引用形如 `mac(%mac1(), variable1, ...)` 的宏

现在我们考虑宏与宏之间嵌套使用的问题,这可能有两种情况:一种是一个完整的宏来引用另一个完整的宏,另一种是将一个宏函数(宏函数的内容将在后面具体介绍)的返回结果作为宏参数来引用。

【例 3-37】 引用定义在宏内部的宏。设程序名为 `SASTJFX3_37.sas`。

```
%macro mac_outside;
%macro mac(var1,var2,var3);
%put &var1 &var2 &var3;
%mend mac;
%mend mac_outside;
%mac_outside;
%mac(data_old1,data_old2,data_old3);
%mac(data_new1,data_new2,data_new3);
run;
```

例中说明了这样一个问题,如果用户要引用定义在宏内部的宏 `mac` 的话,就必须事先引用包含它的宏 `mac_outside`。原因是只有当宏 `mac_outside` 被引用过一次时,宏处理器编译过这个宏内部的程序,宏 `mac` 才被创建。如果去掉 `%mac_outside` 语句的话, `SAS log` 中会给出错误信息,但前提是与 `mac` 同名的宏之前没有被创建过。

【例 3-38】 用宏 `mac_outside` 引用宏 `mac`。设程序名为 `SASTJFX3_38.sas`。

```
%macro mac;
%put &var1 &var2 &var3;
%mend mac;
%macro mac_outside;
%let var1=data1;
%let var2=data2;
%let var3=data3;
%mac;
%mend mac_outside;
%mac_outside;
run;
```

例中定义了两个宏，在定义宏 `mac_outside` 时，我们添加了引用宏 `mac` 的语句 `% mac`。因此，在引用宏 `mac_outside` 时也完成了间接引用宏 `mac` 的操作。例中两个宏的定义顺序可以互相交换，不会影响结果。带参数宏之间的嵌套调用与上面的例子的写法基本相同，下面将程序简单修改如下：

```
%macro mac_outside(var_out1,var_out2,var_out3);
%put &var_out1 &var_out2 &var_out3;
%mac (&var_out1,&var_out2,&var_out3);
%mend mac_outside;
%macro mac(var1,var2,var3);
%put &var1 &var2 &var3;
%mend mac;
%mac_outside(data1,data2,data3);
run;
```

这个例子的运行结果是将 `data1`、`data2`、`data3` 输出了两遍。需要注意的问题是，在定义 `mac_outside` 内部引用宏 `mac` 时括号内参数的写法。如果不加 `&`，就变成了把 `var_out1` 等作为字符串赋给了变量 `var1` 等。若加上 `&`，就变成了把局部宏变量 `var_out1` 等赋给了变量 `var1` 等，要注意区分。

【例 3-39】 引用宏参数是宏函数返回结果的宏。设程序名为 `SASTJFX3_39.sas`。

```
%macro mac(n,var);
%if &n>5
%then % put &var;
%mend mac;
%mac(% length(data),data);
%mac(% length(data_new),data_new);
run;
```

例中创建了一个含有两个参数的宏 `mac`。这个宏选择性地输出字符串，`% length` 函数的功能是返回字符串的长度。也就是说，当这个字符串的长度大于 5，则输出该字符串，否则不进行任何操作，宏 `mac` 选取了 `% length` 的返回结果作为其参数。宏参数在选择上给了用户很大的自由空间，选择一个好的宏参数能够在一定程度上简化程序。

4. 引用含有特殊字符的宏

宏语言的基本单位是字符，变量的值无论是数字的还是字符的都会被统一作为字符来处理，因此宏处理器允许用户在文本中使用各式各样的特殊字符。但同时也衍生出一系列问题。最简单的一个例子：`%` 是作为百分号来使用还是作为调用宏符号来使用。为了避免出现这类混淆，我们必须让宏处理器明确这些字符在当前情形的含义，是作为 SAS 系统设定的一些特殊符号还是仅仅作为文本中的字符。宏引用函数能够根据用户需要屏蔽一些特殊字符，从而使得宏处理器不会错误地解析这些字符，产生所不希望的结果。宏引用时容易造成混淆的特殊字符见表 3-11。

表 3-11 宏引用时容易造成混淆的特殊字符

| | | | | | | | |
|-------|----|-----|----|-----------|---|-----|---|
| blank |) | = | LT | ~ | * | NOT | % |
| ; | (| | GE | , (comma) | / | EQ | & |
| ¬ | + | AND | GT | ' | < | NE | # |
| ^ | -- | OR | IN | " | > | LE | |

一些常见易造成混淆的情况如下。

- % 是作为百分号还是作为宏引用符号。
- 一个文本中不平衡的单引号是不是来自人名。
- & 是表示两事物合并还是表示引用宏变量。
- 逗号是作为参数值还是作为分隔符号。

常用的宏引用函数如下。

- %STR 和%NRSTR：在宏语句中，这两种函数使得宏处理器把特殊字符作为文本来对待。
- %BQUOTE 和%NRBQUOTE：这两种函数使得宏处理器把那些从宏表达式解析过来的特殊字符作为文本对待，该类函数常用在宏或者宏程序语句执行的过程中。
- %SUPERQ：该函数不允许宏处理器对其参数进行任何解析。

(1) %STR 和%NRSTR 函数

当特殊字符影响到宏处理器正常编译宏程序语句时，我们可以使用%STR 和%NRSTR 引用函数来屏蔽掉这些特殊字符，参见表 3-12。

表 3-12 %STR 和%NRSTR 可以屏蔽掉的特殊字符

| | | | | | | | |
|-------|----|-----|----|------------|---|-----|----|
| blank |) | = | NE | ~ | * | OR | GT |
| ; | (| | LE | , (comma) | / | NOT | |
| ¬ | + | # | LT | ‘ | < | IN | |
| ^ | -- | AND | GE | “ | > | EQ | |

%NRSTR 还可以屏蔽掉 & 和%。

注：如果用这两个函数来屏蔽不匹配的单引号、双引号以及括号，需要在这些特殊字符前加一个百分号%。

【例 3-40】 屏蔽不匹配的圆括号。设程序名为 SASTJFX3_40. sas。

```
%let mac_var=%str(exp%(6);
%put &mac_var;
run;
```

第二个圆括号“(”前面的%一定不能忽略，否则会出错，圆括号可以被替换为单引号、双引号，屏蔽方法都是相同的。

一般情况下，我们使用%NRSTR 来屏蔽百分号%。在这样一种情况下，我们可以使用%STR 来屏蔽%：%后面没有会被宏处理器解释为宏名称的文本。与上面的例子类似，%前面需要添加一个%来辅助屏蔽。

【例 3-41】 屏蔽%的特殊情况。设程序名为 SASTJFX3_41. sas。

```
%let mac_var=%str(% % % '% % % ');
%put &mac_var;
run;
```

输出结果：% '% '。从结果不难理解宏处理器的解析原理，它需要%来辅助屏蔽每一个特殊字符的含义，因此上面的变量值中有 4 个%是作为在特殊字符前添加的百分号。

【例 3-42】 文本中含有“；”的情况。设程序名为 SASTJFX3_42. sas。

```
%let mac_var=%str(proc ttest;run;);
%put &mac_var;
%put %str(good done;);
run;
```

用%STR(或%NRSTR)来帮助宏处理器识别出正确的宏程序语句的结尾,如果不使用%STR函数,ttest后面的“;”会被认为是%LET语句的结尾,使得宏变量最终的存储值为proc ttest,无法得到用户想要的结果。

考察第二个%put语句,使用引用函数%STR后,第二个“;”被识别为语句的结尾,得到输出语句为“good done;”,若去掉引用函数,第一个“;”被识别为语句的结尾,得到的输出语句为“good done”。

【例3-43】 输出&+宏变量名称,而非输出宏变量值。设程序名为SASTJFX3_43.sas。

```
%macro mac;
%let mac_var=saber;
%put %nrstr(You get &mac_var!);
%put You get &mac_var!;
%mend mac;
%mac;
run;
```

运行结果为:

You get &mac_var!

You get saber!

在引用函数的辅助下,第一个%put语句中&作为引用宏变量符号的含义被屏蔽掉,而仅仅是作为普通的文本内容输出出来,第二个%put语句中宏处理器顺利将mac_var识别为宏变量而解析出来。当要屏蔽%作为引用宏的符号这一含义时,同样需要使用%NRSTR函数。

看下面这三行语句:

```
%let mac_var=%str(proc ttest;run;);
%let mac_var=proc %str(ttest;)%str(run;);
%let mac_var=proc ttest %str(;)run %str(;);
```

这三行赋值语句的结果是相同的,因为宏处理器只是不解析%STR函数屏蔽掉的特殊字符,而其他文本内容保持原样。因此,第三个%LET语句是最简化的引用方式,但我们并不提倡此举,类似第一个%LET语句的用法,将大块的文本包含在引用函数括号内实际上对宏处理器没有任何危害,而且可以增强程序的可读性。

(2)%BQUOTE和%NRBQUOTE函数

%BQUOTE和%NRBQUOTE在宏或者宏语句执行时屏蔽掉特殊字符。这两个函数可以直接屏蔽掉%STR和%NRSTR能够屏蔽的所有特殊字符,也就是说对于不匹配的单引号、双引号以及圆括号同样可以直接屏蔽,不用像%STR和%NRSTR函数额外在前面添加一个百分号%。此外,这两个函数在屏蔽之前,能够发现那些无法引用的宏变量或者宏,并给出警告信息。

%BQUOTE函数执行时将宏变量或宏解析时产生的所有圆括号和引号都看成特殊字符,因此,该函数不会介意引号或者圆括号个数是否匹配。

%NRBQUOTE函数对于解析第一次遇到的宏变量值很有用,它能够阻止%EVAL函数将&和%解析成操作符。如果%NRBQUOTE的参数中含有无法解析的宏变量引用或者宏调用,在它将与%屏蔽之前会给出相应的警告信息。

由于%BQUOTE和%NRBQUOTE函数是在执行期间操作的,并且比%STR和%NRSTR函数功能更灵活,因此使用%BQUOTE和%NRBQUOTE函数来屏蔽包含宏变量引用文本中的特殊字符是很好的选择。

【例 3-44】 用法举例。设程序名为 SASTJFX3_44.sas。

```
data test;
store = "Mr'Smith";
call symput('mac_var',store);
run;
%put %bquote(&mac_var);
```

例中介绍了% BQUOTE 函数的一个简单用法，在数据步中将 Mr'Smith 分配给 store，需要将字符串包含在双引号中，这样才可以使用不匹配的单引号，CALL SYMPUT 将 store 存储的字符串作为宏变量值赋给 mac_var。在% PUT 语句中，如果不使用 &BQUOTE 函数，宏处理器将给出错误信息提示% PUT 中有无效的操作符。

(3) % SUPERQ 函数

% SUPERQ 函数不允许对其参数值进行任何解析，在宏运行期间能够将所有需要屏蔽的特殊文本屏蔽，宏处理器不会给出任何关于宏变量调用和宏引用的警告信息，因为% SUPERQ 函数阻止了宏处理器对其参数的解析。因此，即便% NRBQUOTE 函数能够使程序正常工作，我们也倾向于选择% SUPERQ 函数来避免一些不想要的警告信息。

% SUPERQ 函数从宏符号表中获得宏变量值之后，立即进行引用，同时阻止宏处理器对其函数参数做进一步解析。例如，一个宏变量 mac_var 被解析为 Tom&Jerry，使用% SUPERQ 函数会阻止宏处理器对 &Jerry 做进一步解析。

【例 3-45】 使用% SUPERQ 函数阻止宏处理器对其参数进行解析。设程序名为 SAS-TJFX3_45.sas。

```
%window ask
#5 @5 'Enter two values:'
#5 @24 mac_var 15 attr=underline;
%macro mac;
%put *** Jack Jones.*** ;
%mend;
%macro test;
%display ask;
%put *** %superq(mac_var)*** ;
%mend;
%test;
```

我们在命令行中输入% var % mac，然后得到了这样的结果：*** % var % mac ***，SAS log 中也没有给出任何警告信息。以上结果说明宏 mac 没有被调用过，因此 mac 中的% put 语句也没有被执行，var 这个宏实际上没有被定义过，但 SAS log 中也没有警告信息，说明% SUPERQ 函数阻止了宏处理器对% mac % var 的进一步解析。

不同的引用函数能够屏蔽不同的特殊字符，表 3-13 简单总结了各引用函数能够屏蔽的特殊字符。

表 3-13 各引用函数能够屏蔽的特殊字符

| 特殊字符 | 宏引用函数 |
|---|---|
| + -- * / < > = ~ ^ ~ ; , # blank AND OR NOT EQ NE LE LT GE GT IN | 所有引用函数均可屏蔽 |
| &% | % NRSTR, % NRBQUOTE, % SUPERQ, % NRQUOTE |
| 不匹配的'“() | % BQUOTE, % NRBQUOTE, % SUPERQ, % STR *, % NRSTR *, % QUOTE *, % NRQUOTE * |

3.2.5 常用宏语句和系统宏函数

1. 宏表达式

宏表达式大致分为三类：文本、逻辑和算术表达式。文本表达式是由文本、宏变量、宏函数以及宏引用任意组合而成的，经过宏处理器的解析，表达式转化为生成文本。逻辑表达式和算术表达式都由一系列运算符和运算对象组合而成，并能根据这个表达式求出一个结果。逻辑表达式使用逻辑运算符，算术表达式使用算术运算符。

(1) 逻辑表达式和算术表达式的定义

操作对象：算术表达式和逻辑表达式中的操作对象必须是文本，当对表达式求值时，表示数字的操作对象会被临时转化成算术数值参与计算。宏处理器默认使用整数算法，只有整数和十六进制数值才能进行算术运算，带有小数点的数字文本是不会被转化为数值参与计算的。利用%SYSEVAL 函数可以将含有小数点的数字文本临时转化成浮点数进行运算。

操作符：宏表达式的操作符是数据步中操作符的子集，如宏语言中没有 MAX 和 MIN 操作符，并且不能识别“：”。宏语言表达式中各运算操作执行的顺序与数据步相同，括号内的操作优先级最高。

在一些特定的宏函数(如% EVAL)或者宏语句(如% DO % WHILE)中都可以使用逻辑表达式和算术表达式，通过这些表达式可以控制宏函数和宏语句产生所需的文本。

也可以使用文本表达式来产生部分或者完整的算术以及逻辑表达式。宏处理器在求表达式值之前会先对表达式进行解析。

【例 3-46】 利用文本表达式产生算术表达式。设程序名为 SASTJFX3_46. sas。

```
%let x=3;
%let y=4;
%let operator=* ;
%put &x &operator &y=%eval(&x &operator &y);
```

输出结果为：3 * 4 =12。可以看到，在% EVAL 函数中，宏处理器先解析了三个宏变量 x、y、operator，然后计算了表达式 3 * 4 的结果。

(2) 宏处理器如何处理算术表达式

宏设备是一个字符串操作设备，不过，在某些特殊的情形下，宏处理器可以计算表示数值的操作对象。默认宏处理器只能完成整数的算术计算，利用% SYSEVALF 函数可以完成浮点数的运算。下面举例说明在宏语句中进行整数算术操作。

【例 3-47】 利用宏语句进行整数算术操作。设程序名为 SASTJFX3_47. sas。

```
%let x=%eval(4+7);
%let y=%eval(8*5);
%let p=%eval(6/2);
%let q=%eval(8/3);
%put The result of x is &X;
%put The result of y is &Y;
%put The result of p is &P;
%put The result of q is &Q;
```

运行结果：

The result of x is 11

```
The result of y is 40
The result of p is 3
The result of q is 2
```

我们注意到 q 值的计算结果只保留了整数部分，也就是说 %EVAL 函数完成除法操作时，会将小数部分舍弃。

由于 %EVAL 函数只支持整数算术操作，当宏处理器遇到无法转换成整数数值的字符时（如浮点数），宏处理器会给出错误信息提示。下面举例说明在宏语句中进行浮点数算术操作。

【例 3-48】 利用宏语句进行浮点数算术操作。设程序名为 SASTJFX3_48.sas。

```
%let x=%sysevalf(4.3+7.2);
%let y=%sysevalf(8.4*5.11);
%let p=%sysevalf(6.0/2.0);
%let q=%sysevalf(8.0/3.0);
%put The result of x is &X;
%put The result of y is &Y;
%put The result of p is &P;
%put The result of q is &Q;
```

运行结果：

```
The result of x is 11.5
The result of y is 42.924
The result of p is 3
The result of q is 2.666666666666666
```

当 %SYSEVALF 函数计算算术表达式时，表示数值的操作对象会被临时转换成浮点数，并且得到的结果也是浮点数。%SYSEVALF 函数还提供了数据类型转换功能，为的是与其他宏表达式能够更好地兼容。

【例 3-49】 利用 %SYSEVALF 函数转换数据类型。设程序名为 SASTJFX3_49.sas。

```
%let x=3.5;
%put %sysevalf(&x,boolean);
%put %sysevalf(&x,integer);
%put %sysevalf(&x,ceil);
%put %sysevalf(&x,floor);
```

得到的结果分别为 1、3、4、3，其中 integer 是只保留整数部分，ceil 是对上取整，floor 是对下取整。

(3) 宏处理器如何处理逻辑表达式

逻辑表达式的返回值只有 true 和 false 两种结果，在宏语言中，任何非零的数值都返回 true，而只有 0 值返回 false。

当宏处理器计算逻辑表达式时，表示数字的字符会被临时转换为数值。

【例 3-50】 逻辑表达式中的整数比较操作。设程序名为 SASTJFX3_50.sas。

```
%macro compare(x,y);
%if &x>&y %then %put &x is greater than &y;
%else %if &x=&y %then %put &x equals &y;
%else %put &x is less than &y;
%mend compare;
%compare(2,1);
%compare(0,0);
%compare(-1,1);
```

运行结果:

2 is greater than 1
0 equals 0
- 1 is less than 1

从上面的运行结果可以看出, 逻辑表达式中表示数字的操作对象被临时转换为数值进行比较。包含浮点数和缺失值的逻辑表达式, 要借助%SYSEVALF 函数完成比较操作。

【例 3-51】 逻辑表达式中的浮点数比较操作。设程序名为 SASTJFX3_51.sas。

```
%macro compare(x,y);  
%if %sysevalf(&x>&y) %then %put &x is greater than &y;  
%else %if %sysevalf(&x=&y) %then %put &x equals &y;  
%else %put &x is less than &y;  
%mend compare;  
%compare(2.1,1.1);  
%compare(.,0);  
%compare(-1.1,1.2);
```

运行结果:

2.1 is greater than 1.1
. is less than 0
- 1.1 is less than 1.2

需要注意这样一种情况: 有些操作对象看起来是整数却包含小数点(如 2.0)。这时, 如果不借助%SYSEVALF 函数, 宏处理器就会将其视为字符文本, 并进行字符值的比较, 就会出现 10 比 2.0 小这样意想不到的结果。实际应用中要避免这种问题的出现。

2. 常用宏语句

宏语句通常由关键字字符串、SAS 名、特殊字符和操作对象组合而成, 以分号作为结束, 用户通过编写宏语句来指导宏处理器执行一项操作。一部分宏语句只能用在宏定义中, 另一部分宏语句可以用在整个 SAS 程序的任何地方, 参见表 3-14 和表 3-15。

表 3-14 在宏定义内外均可以使用的宏语句

| 语 句 | 功能描述 |
|-----------|----------------------------|
| % * | 指定注释内容 |
| % COPY | 从 SAS 库中复制规定的内容 |
| % DISPLAY | 显示宏窗口 |
| % GLOBAL | 创建全局变量 |
| % INPUT | 宏执行时给宏变量赋值 |
| % LET | 创建宏变量或者给宏变量赋值 |
| % MACRO | 开始宏定义 |
| % PUT | 把文本或者宏变量值输出到 SAS log 窗口中 |
| % SYMDEL | 删除指定的在参数中命名的宏变量 |
| % SYSCALL | 调用 SAS call 程序 |
| % SYSEXEC | 发出操作系统指令 |
| % SYSLPUT | 在远程主机上定义新的宏变量或者更改已经存在的宏变量值 |
| % SYSRPUT | 把远程主机上宏变量值赋给本地主机上的宏变量 |
| % WINDOW | 自定义用户窗口 |

表 3-15 只有在宏定义内可以使用的宏语句

| 语 句 | 功能描述 |
|--------------------|-------------------|
| % ABORT | 终止伴随 SAS 程序正在执行的宏 |
| % DO | 开始% DO 语句组 |
| 循环% DO | 根据索引变量值循环执行语句 |
| % DO % UNTIL | 重复执行语句直到条件为真 |
| % DO % WHILE | 当一个条件为真时重复执行语句 |
| % END | 结束% DO 语句组 |
| % GOTO | 使宏处理器跳转到指定的标签 |
| % IF-% THEN/% ELSE | 有条件地执行宏 |
| % label: | 定义% GOTO 的目标语句 |
| % LOCAL | 创建局部宏变量 |
| % MEND | 结束宏定义 |
| % RETURN | 正常结束当前执行宏 |

上面部分语句的功能已经在前面的例子多次用到(如%let, %macro, %put等),用户可以回顾上文中的内容来熟悉这些语句的功能。下面着重介绍上文中没有提到的且在实际应用中很常见的宏语句。

(1) %DO 语句

该语句使用格式如下:

```
%DO;  
文本和各种宏语句;  
%END;
```

%DO 语句作为宏定义内部一块程序单元的开始,以%END 作为结束,这个单元块就是一个%DO 语句组。%DO 语句组可以互相嵌套。%DO 语句常与%IF-%THEN%ELSE 语句连用,用来指定当%IF 条件满足时宏处理器要执行的程序模块。

【例 3-52】 %DO 语句与%IF-%THEN%ELSE 语句连用。设程序名为 SASTJFX3_52. sas。

```
%macro mac(n,var);  
%if &n>5 %then  
%do;  
    %put The variable is &var;  
    %put The length of the variable is greater than 5;  
%end;  
%else  
%do;  
    %put The variable is &var;  
    %put The length of the variable is not greater than 5;  
%end;  
%mend;  
%mac(%length(data_new),data_new);  
%mac(%length(data),data);
```

%IF 和%ELSE 根据参数 VAR 的字符串长度不同选择执行不同的%DO 组语句。当一个条件下需要执行多个语句(比如执行一个过程)时,我们就可以把这些语句都写到%DO 组内部。

(2) 循环%DO 语句

该语句使用格式如下:

```
%DO MAC_VAR = 开始值 %TO 终止值 <%BY 增量>;  
文本和各种宏语句;  
%END;
```

其中,MAC_VAR 也可以是由文本表达式产生的,开始值与终止值决定了%DO 语句的循环次数。如果 MAC_VAR 并不存在于宏符号表中,那么宏处理器会临时创建 MAC_VAR 并将它置于局部符号表内。增量决定了每经过一次循环后 MAC_VAR 的改变幅度,默认值是 1,选用合适的增量便于用户控制%DO 的循环次数。

【例 3-53】 批量创建宏变量。设程序名为 SASTJFX3_53. sas。

```
%macro create(number);  
%do i = &number %to1 %by-2;  
    %let var&i = &i;  
    %put var&i = &&var&i;
```

```
%end;
%mend;
%create(11);
```

运行结果:

```
var11 = 11
var9 = 9
var7 = 7
var5 = 5
var3 = 3
var1 = 1
```

利用循环% DO 语句可以很方便地创建规则类似且数目较多的宏变量。

(3) % DO % UNTIL 语句

该语句使用格式如下:

```
%DO %UNTIL(表达式);
```

文本和各种宏语句;

```
%END;
```

其中,表达式可以是任何能产生合理数值的表达式,宏处理器会在每次循环最后计算表达式的结果。只有当结果为0时,该表达式才为假,其余情况均为真。如果表达式返回了一个无效的结果或者结果含有非数值的字符,那么宏处理器会给出错误信息。

由于% DO % UNTIL 语句是在每次循环最后检查表达式结果是否为真,也就是说% DO % UNTIL 内的语句会至少执行一次。

【例 3-54】 寻找某个字母在某单词中出现次数。设程序名为 SASTJFX3_54. sas。

```
%macro mac(letter,string);
%let i=1;
%let count=0;
%do %until(&i > %length(&string));
%if %substr(&string,&i,1) = &letter %then
%do;
%let i = %eval(&i + 1);
%let count = %eval(&count + 1);
%end;
%else
%let i = %eval(&i + 1);
%if &i = %length(&string) + 1 %then
%if &count = 0 %then
%put There is no &letter in &string..;
%else
%put The &string has &count &letter..;
%end;
%mend;
%mac(s,statements);
%mac(t,statements);
%mac(c,statements);
```

运行结果:

The statements has 2 s.
 The statements has 3 t.
 There is no c in statements.

用宏变量 count 来统计规定字母的出现次数,用% SUBSTR 函数来扫描单词的每个位置。如果出现了规定的字母,则 count 加 1。如此循环,直到单词的所有位置均被扫描了一遍,最后输出结果。

(4) % DO % WHILE 语句

该语句使用格式如下:

```
%DO %WHILE(表达式);
  文本和各种宏语句;
%END;
```

其中,表达式可以是任何能产生合理数值的表达式,宏处理器会在每次循环的开始计算表达式的结果,只有当结果为 0 时该表达式才为假,其余情况均为真,如果表达式返回了一个无效的结果或者是结果含有非数值的字符,那么宏处理器会给出错误信息。

由于% DO % UNTIL 语句是在每次循环最初检查表达式结果是否为真,也就是说如果条件为假,% DO % UNTIL 内的语句将会一次都不执行。

【例 3-55】 寻找某个字母在某单词中第一次出现的位置。设程序名为 SASTJFX3_55. sas。

```
%macro mac(letter,string);
%let i=1;
%let flag=0;
%do %while(&i <=%length(&string) and(&flag=0));
%if %substr(&string,&i,1) = &letter %then
  %do;
  %let flag=1;
  %put The first position of &letter in &string is &i..;
%end;
%else
  %let i=%eval(&i+1);
%if(&i-1) = %length(&string) %then
  %put There is no &letter in &string..;
%end;
%mend;
%mac(m,macro);
%mac(e,statement);
%mac(t,backspace);
```

运行结果:

The first position of m in macro is 1.
 The first position of e in statement is 5.
 There is no t in backspace.

FLAG 是标志值,用来标志要寻找的字母是否在该单词中已经出现。若出现,则立即将 FLAG 置为 1 以跳出循环。关于% SUBSTR 和% LENGTH 等函数将在后面进行介绍。

(5) % GOTO 语句和% LABEL 语句

该语句使用格式如下:

```
%GOTO label;
%label: 宏文本;
```

用%GOTO 跳转到分支语句有两个限制：其一，%GOTO 跳转的目标语句必须存在于当前宏，该语句无法完成从一个宏跳转到另一个宏的操作；其二，如果%DO 循环语句、%DO %UNTIL 语句或%DO %WHILE 语句当前并没有执行，即使利用%GOTO 语句跳转到这类语句内部也不能引起它们的运行。

在定义 label 名时，前面需要使用百分号，而在%GOTO 语句中声明 label 时，无须使用百分号。

【例 3-56】 简单的语句跳转实现。设程序名为 SASTJFX3_56.sas。

```
%macro mac(var);
  %if %length(&var) <=5 %then %goto fin;
  %let var=%substr(&var,1,5);
  %fin:%put var=&var;
%mend;
%mac(phone);
%mac(telephone);
```

运行结果：

var = phone

var = telep

该程序完成了这样一个操作：如果输入的字符串长度不超过 5，那么跳转到 FIN 直接输出字符串；如果输入的字符串长度超过 5，那么截取该字符串前 5 个字符再将其输出。

(6) %IF-%THEN/%ELSE 语句

该语句使用格式如下：

```
%IF 表达式 %THEN 操作;
<%ELSE 操作; >
```

%IF-%THEN/%ELSE 语句在前面的举例中已经多次用到，这里不再单独举例。下面简单说明%IF-%THEN/%ELSE 语句和 IF-THEN/ELSE 语句的区别。本质上，两种语句是属于不同的语句体系的，%IF-%THEN/%ELSE 语句属于宏语言范畴，功能是有条件地生成文本，而 IF-THEN/ELSE 语句属于 SAS 基本语言范畴，功能仅限于在数据步中有条件地完成某项操作；%IF-%THEN/%ELSE 语句中的条件表达式包含恒定文本或者是能够产生文本的文本表达式，而 IF-THEN/ELSE 语句中的条件表达式只能是数据步变量、字符常量、数值常量或者是数据和时间常量；如果数据步中的部分文本是由%IF-%THEN/%ELSE 语句产生的，那么这些文本将被数据步编译器编译并执行。

(7) %RETURN 语句

%RETURN 语句可以正常终止当前宏的运行。

【例 3-57】 利用%RETURN 正常结束宏的运行。设程序名为 SASTJFX3_57.sas。

```
%macro test(error);
  %if &error=1 %then %return;
  data a;
    x=1;
  run;
%mend;
%test(0);
%test(1);
```

当参数值为 0 时,宏 TEST 顺利建立了数据集 A,当参数值为 1 时,由于此时满足 % IF 的条件,宏将由于 % RETURN 语句而结束运行。

3. 常用系统宏函数

宏语言中的函数能够处理一个或多个参数并产生相应的结果,它将宏和宏参数很好地结合起来,更便于用户根据需要随时进行引用。宏函数种类繁多,SAS 系统为用户设计了很多宏函数实现了各式各样的功能,参见表 3-16。

表 3-16 宏函数分类

| 函数种类 | 函数名称 | 函数功能 |
|-----------------------|-----------------------|--|
| 宏计算函数 (详见 3.2.1 节) | % EVAL | 计算算术和逻辑表达式(整数格式) |
| | % SYSEVALF | 计算算术和逻辑表达式(浮点数格式) |
| 宏字符函数 | % INDEX | 返回某字符串在文本中第一次出现的位置 |
| | % LENGTH | 返回一个字符串的长度 |
| | % SCAN, % QSCAN | 扫描单词,后者可以扫描包含 % 和 & 的单词 |
| | % SUBSTR, % QSUBSTR | 提取指定的字符串,后者可以提取包含 % 和 & 的字符串 |
| | % UPCASE, % QUPCASE | 转换小写字符为大写字符,后者可以转换包含 % 和 & 的小写字符为大写 |
| 宏引用函数 (详见 3.2.4 节) | % BQUOTE, % NRBQUOTE | 宏执行时屏蔽解析值内的各种特殊字符,在屏蔽不匹配的引号或者括号时不需要被标记 |
| | % QUOTE, % NRQUOTE | 宏执行时屏蔽解析值内的各种特殊字符,在屏蔽不匹配的引号或者括号时需要 % 作为标记 |
| | % STR, % NRSTR | 宏编译时屏蔽恒定文本内的各种特殊字符,在屏蔽不匹配的引号或者括号时需要 % 作为标记 |
| | % SUPERQ | 宏执行时屏蔽所有特殊字符,阻止宏处理器的任何解析 |
| | % UNQUOTE | 对于某个指定值不屏蔽任何特殊字符 |
| 其他宏函数 | % SYMEXIST | 返回指定的宏变量是否存在 |
| | % SYMGLOBAL | 返回指定的宏变量是否是全局宏变量 |
| | % SYMLocal | 返回指定的宏变量是否是局部宏变量 |
| | % SYSFUNC, % QSYSFUNC | 在宏设备中执行 SAS 函数或者自定义函数 |
| | % SYSGET | 返回指定的主机环境变量值 |
| | % SYSPROD | 返回 SAS 软件产品的许可性 |

在前几节的介绍中,我们已经多次提到并运用了部分函数,这里再进一步说明几个常用的 SAS 宏函数。

(1) % INDEX 函数

该函数使用格式如下:

```
%INDEX(source,string)
```

source 和 string 都是来自于文本中的字符串。这个函数将从 source 中查找 string 第一次出现的位置,并以数字形式返回 string 首字母所处的位置,如果没有查找到则返回 0。

【例 3-58】 利用 % INDEX 查询字符或者字符串。设程序名为 SASTJFX3_58.sas。

```
%let var = to see a world in a grain of sand;
%put %index(&var,a);
%put %index(&var,or);
%put %index(&var,ab);
```

运行结果: 8 11 0

读者可以自行验证运行结果来熟悉% INDEX 函数的功能。

(2)% SCAN, % QSCAN 函数

该函数使用格式如下:

```
%SCAN(var,n<,>,delimiters>)
%QSCAN(var,n<,>,delimiters>)
```

var 是字符串或者是文本表达式, 如果 var 包含特殊字符, 需要使用% QSCAN 函数。如果 var 包含逗号, 需要借助宏引用函数% QUOTE 来引用这个参数 var。

n 是一个整数或者是可以得到整数结果的表达式, 它指定了函数返回的字符位置, 如果 n 超过了参数 var 的长度, 那么函数将返回一个空的字符串。

在 n 之后还可以定义参数 var 中的分隔符作为单词之间的区分, 由于% SCAN 函数默认很多字符作为它的分隔符。如果只需要空格或者逗号作为其分隔符, 则需要在 n 之后补充一个参数% STR()或者% STR(,), 这时其他的字符就被视为文本了。

常见的默认分隔符: blank . < (+ & ! \$ *) ; ^ - / , % |

【例 3-59】 % SCAN 和% QSCAN 函数的对比和使用。设程序名为 SASTJFX3_59. sas。

```
%let var=welcome to|China;
%put %scan(&var,2);
%put %scan(&var,2,%str( ));
%let var1=%nrstr(The answer is &var);
%put %scan(&var1,4);
%put %scan(&var1,4,%str( ));
%put %qscan(&var1,4,%str( ));
```

运行结果:

```
to
to|China
var
welcome to|China
&var
```

% SCAN 函数默认空格和“|”都是分隔符, 因此第一个% PUT 只输出 to, 而在第二个% PUT 中, 我们规定了只有空格才能作为分隔符, 因此“|”被视为文本; 第三个% PUT 与第四个% PUT 与前面的情况相同, 而% QSCAN 函数屏蔽了“&”符号, 因此直接输出 &VAR, 而不是宏变量 VAR 的解析结果。

% QSCAN 函数可屏蔽的特殊字符如下: &% ' "() + - * / < > = ˆ ~ ; , # blank AND OR NOT EQ NE LE LT GE GT IN

(3)% SUBSTR, % QSUBSTR 函数

该函数使用格式如下:

```
%SUBSTR(var,n<,>,length>)
%QSUBSTR(var,n<,>,length>)
```

var 是字符串或者是文本表达式, 如果 var 包含特殊字符, 则需要使用% QSUBSTR 函数。

n 是一个整数或者是可以得到整数结果的表达式, 它指定了亚字符串首字母的位置。如果 n 超过了参数 var 的长度, 那么函数将给出警告信息并返回一个空值。

LENGTH 是一个可选的参数来指定亚字符串的字符数量。如果 LENGTH 超过了 n 之后余下的字符串长度，那么函数将给出警告信息并返回从 n 开始到末尾的所有字符。如果未定义参数 LENGTH，默认返回的就是从 n 到末尾的所有字符。

【例 3-60】 %SUBSTR 和 %QSUBSTR 函数的对比和使用。设程序名为 SASTJFX3_60.sas。

```
%let a=one;
%let b=two;
%let c=three;
%let var=%nrstr(&a&b&c);
%put %substr(&var,2,1);
%put %substr(&var,2,2);
%put %substr(&var,2,3);
%put %qsubstr(&var,2,3);
```

运行结果：

```
a
a&
atwo
a&b
```

可以看到，%SUBSTR 函数根据亚字符串首字母的位置和规定的长度进行扫描。如果扫描到解析符号和已存在的宏变量，则宏处理器进行解析，通过%QSBSTR 函数可以屏蔽掉解析符号“&”。

%QSUBSTR 函数可屏蔽的特殊字符如下：&% ' "() + - * / < > = ˆ ~ ; , # blank AND OR NOT EQ NE LE LT GE GT IN

(4) %UPCASE, %QUPCASE 函数

该函数使用格式如下：

```
%UPCASE (字符串或文本表达式)
%QUPCASE (字符串或文本表达式)
```

这两个函数实现了参数中小写字母向大写字母的转换，区别在于后者可以屏蔽特殊字符。如果需要实现逆向转换，可以使用%LOWCASE 和%QLOWCASE 函数。

【例 3-61】 %UPCASE 函数和 %QUPCASE 函数的对比和使用。设程序名为 SASTJFX3_61.sas。

```
%let a=sas;
%let b=%nrstr(&a);
%put %upcase(&b);
%put %upcase(%upcase(&b));
%put %qupcase(&b);
```

运行结果：

```
sas
SAS
&A
```

实际应用中需要大小写转换的场合有很多，这两个函数需要灵活掌握。

%QUPCASE 函数可屏蔽的特殊字符如下：&% ' "() + - * / < > = ˆ ~ ; , # blank AND OR NOT EQ NE LE LT GE GT IN

3.2.6 宏与其他模块接口

1. 宏与数据步接口

宏与数据步之间的接口由 8 种工具组成。通过合理地运用这些工具，可以在数据步过程中影响到宏设备的运作。通常，宏设备的工作是发生在数据步之前的，也就是说在数据步执行时，宏语句提供的信息已经被处理过了。用户通过使用接口可以实现如下几种操作：

- 在 SAS 程序中实现数据步及其相邻下一步之间的信息传递；
- 根据数据步执行信息来调用一个宏；
- 在数据步中解析宏变量；
- 删除宏变量；
- 从宏设备中传递宏变量信息给数据步。

表 3-17 列出了宏与数据步接口。

表 3-17 宏与数据步接口

| 分 类 | 工 具 | 功能描述 |
|-----|------------------|--------------------------------|
| 执行 | CALL EXECUTE 子程序 | 对参数解析，并立即执行解析值或者在数据步执行完成后执行解析值 |
| 解析 | RESOLVE 函数 | 在数据步中解析文本表达式 |
| 删除 | CALL SYMDEL 子程序 | 删除参数中指定的宏变量 |
| 信息 | SYMEXIST 函数 | 返回宏变量是否存在的指示信息 |
| 读/写 | SYMGET 函数 | 在数据步中返回宏变量的值 |
| 信息 | SYMGLOBL 函数 | 返回宏变量是否是全局的指示信息 |
| 信息 | SYMLOCAL 函数 | 返回宏变量是否是局部的指示信息 |
| 读/写 | CALL SYMPUT 子程序 | 将数据步中产生的值赋给宏变量 |

读者可参考 SAS 说明书来了解每个函数或者子程序的具体使用方法，这里仅以 RESOLVE 函数为例来说明数据步接口功能的必要性。

RESOLVE 函数的用法并不复杂，函数只涉及一个参数，但参数的来源可以有很多。

- 可以用单引号标注的文本表达式。当参数涉及宏变量引用时，RESOLVE 函数会尝试去解析该引用。如果该宏变量不存在，RESOLVE 函数直接返回一个未解析的引用。

```
Var = resolve('%mac');  
Var = resolve('&mac_var');
```

- 可以是变量值是文本表达式的数据步变量。

```
Addr = '&mac_var';  
Var = resolve(addr);
```

- 可以是能够产生宏设备可解析文本的字符表达式。

```
Var = resolve('%mac' || left(name));
```

【例 3-62】 利用 RESOLVE 函数解析宏或宏变量引用。设程序名为 SASTJFX3_62.sas。

```
%let var = one;  
%macro num;  
  two three  
%mend;  
data test;  
  length var1 - var3 $15;  
  when = '%num';
```

```
var1 = resolve('&var');
var2 = resolve('%num');
var3 = resolve(when);
put var1 = var2 = var3 =;
run;
```

运行结果：

```
var1 = one var2 = two three var3 = two three
```

VAR2 和 VAR3 最后得到相同的结果，但在参数选取上却是用了两种不同方法：一种是直接给出文本表达式，另一种是用取值为文本表达式的数据步变量。

2. 宏与 SQL 接口

SQL 在数据库检索和更新方面的应用是十分广泛的，SAS 软件的 SQL 处理器能够完成以下几种操作：

- 创建表格与视图；
- 检索表格中存储的数据；
- 检索 SQL 以及 SAS/ACCESS 视图中存储的数据；
- 添加或修改表格中的数值；
- 添加或修改 SQL 以及 SAS/ACCESS 视图中存储的数据。

SQL 在 SELECT 语句中为用户提供了创建宏变量的 INTO 子句，并且宏变量的种类也是可选的，INTO 子句和 % LET 语句有着一致的范围规则。

通过使用 SQL 过程提供的宏工具，还可以根据数据有条件地执行程序以及当错误发生时终止程序。表 3-18 给出了 SQL 创建的宏变量的相关功能。

表 3-18 SQL 创建的宏变量的相关功能

| 宏 变 量 | 功能描述 |
|-------------|---|
| SQLEXITCODE | 包含了某些 SQL 模式插入失败时所生成的返回代码，当 SQL 过程结束后，这些返回代码都被写入了宏变量 SYSERR |
| SQLQBS | 包含了 SELECT 语句产生的行数或者观测数 |
| SQLQOPS | 包含了 SQL 过程内部循环处理的重复次数 |
| SQLRC | 包含了 SQL 语句的返回代码 |
| SQLXMSG | 包含了 Pass-Through 设备返回的指示错误的相关代码和描述信息 |
| SQLXRC | 包含了由 DBMS 指定的 Pass-Through 设备的返回代码 |

更多详细信息可以参考 SAS 说明书。

3. 用户自定义宏的存储

当用户提交了一个宏定义之后，宏处理器进行编译，然后默认将其存储到 SAS 的 WORK 库中，这些宏只有在 SAS 软件运行期间才会存在。为了能够存储一些可能会频繁利用的宏，我们需要使用 AUTOCALL 宏设备或者存储编译宏设备。

AUTOCALL 宏设备将 SAS 宏存在一个名为 AUTOCALL 库的外部文件中，利用 AUTOCALL 设备可以很容易地建立这样一个存储位置，便于不同的程序或者不同的用户进行对宏的调用。AUTOCALL 库之间可以进行合并。

3.3 SAS SQL 介绍

3.3.1 SQL 简介

SQL 译为结构化查询语言，是英文 Structured Query Language 的缩写。SQL 广泛应用于关系数据库和表的数据检索更新操作中。

所谓关系数据库，是一种基于关系模型建立起的数据库。相对于关系模型，还有层次模型、网状模型、面向对象模型等。这些模型是建立其相应类型数据库的基础，是这些数据库存储组织数据的基本逻辑。非关系模型的数据库系统在 20 世纪 70 年代到 80 年代初非常流行，在数据库产品中占主体地位，现在已经逐渐被关系模型的数据库系统所取代。关系模型最早是由 IBM 研究员 E. F. Codd 提出的，SQL 也是基于这一模型建立起的一套标准的查询语句。

关系模型是建立在严格的数学概念基础上的。简单来说，从用户观点来看，关系模型由一组关系组成。每个关系的数据结构是一张规范的二维表。这里的表类似与 SAS 中的数据集，每个数据集代表一个关系。数据库中的每一行称为一个元组 (Tuple)，对应 SAS 的观测 (observation)。每一列称为属性，对应 SAS 中的每一个变量。

五元关系表如表 3-19 所示，每行元素称为一个五元组。

通常，我们将一个关系记为：关系名 (属性 1, 属性 2, 属性 3, ..., 属性 n)。

若把上表关系名记为 R，则表 3-19 所代表的关系可记为 R (Name, Sex, Age, Height, Weight)。

例如，在一个实际的学生管理系统应用中，可以存在多个关系：

学生(学号, 姓名, 年龄, 性别, 系别, 年级)

课程(课程号, 课程名, 学分)

选修(学号, 课程号, 成绩)

通过上面的三个关系，如果希望查询某个同学选修的所有的课，从逻辑上看，首先需要学生关系表，通过姓名查询学号；然后在选修关系表中通过学号查询所有这个同学选修的课号；最后在课程关系表中，查询所有课号的名称并输出。之所以我们为学生管理系统不厌其烦地建立多张表，是为了大大方便以后的管理和操作。如果我们将这些信息都列在一起，那么数据的复杂度、存储空间会大大增加，共享性大大降低。

另外需要注意的是，在关系中，一般都存在一个或多个主属性，或称为主码，这个属性在其所在的关系表中是唯一的。例如上面提到的“学生”关系中的“学号”，“选修”关系中的“学号 + 课程号”。在 SAS 中，我们平时所用到的数据集，SAS 会默认赋给其观测号，实际上也相当于其主码。

SQL 语言就是对一个又一个关系表，以及表与表之间进行查询等操作的。

SAS 中的 SQL 过程，是对于标准 SQL 的一个扩展。可以通过 SQL 过程操作任意数据集或外部数据库文件。从功能上来说，SQL 和 DATA 步基本上是一致的，二者各有利弊，用户可根

表 3-19 五元关系表

| Name | Sex | Age | Height | Weight |
|---------|-----|-----|--------|--------|
| Alfred | M | 14 | 69 | 112.5 |
| Alice | F | 13 | 56.5 | 84 |
| Barbara | F | 13 | 65.3 | 98 |
| Carol | F | 14 | 62.8 | 102.5 |
| Henry | M | 14 | 63.5 | 102.5 |
| James | M | 12 | 57.3 | 83 |
| Jane | F | 12 | 59.8 | 84.5 |
| Janet | F | 15 | 62.5 | 112.5 |
| Jeffrey | M | 13 | 62.5 | 84 |

据方便性进行使用。SAS 的 SQL 过程同其他过程一样，对于 SAS 通用的选项、函数、INFORMAT、FORMAT 同样可以使用。

SQL 过程可以实现以下功能：

- 生成报告；
- 生成汇总基本统计信息；
- 从表或视图中检索数据；
- 将表或视图中的数据进行整合；
- 创建表、视图或索引；
- 更新数据；
- 对数据管理系统(Database management system, DBMS)进行操作。

3.3.2 SQL 过程的语句介绍

本节介绍的内容会很多且琐碎，在后面的每一小节都会从功能出发，详细介绍语句。

另外，在 SAS 中，所有的 SQL 语句必须包含在“proc sql”声明后，多个查询语句可以都写在一个声明后，不需要多次声明。

```
proc sql;
/* To Do:在这里写下 SQL 语句* /
quit;
```

例子中使用到数据集 Student 和 Scores，如表 3-20 和表 3-21 所示。

表 3-20 数据表 student(只列出部分)

| StuNum | Subject | Sex | School | PE | EG |
|------------|---------|-----|--------|----|----|
| 1138010104 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1138010108 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1138010114 | 文科 | 女 | 北京八中 | 团员 | 汉族 |
| 1138010128 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1138010126 | 文科 | 男 | 北京八中 | 群众 | 汉族 |
| 1141213504 | 文科 | 女 | 北京六中 | 团员 | 回族 |
| 1138010213 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1241253209 | 理科 | 男 | 北京六中 | 团员 | 汉族 |
| 1138010223 | 文科 | 女 | 北京八中 | 团员 | 汉族 |
| 1138020411 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1141213520 | 文科 | 男 | 北京六中 | 群众 | 汉族 |

表 3-21 数据表 scores(只列出部分)

| StuNum | Math | Chinese | English | Physics | Chemistry | Biology | History | Geography | Politics |
|------------|------|---------|---------|---------|-----------|---------|---------|-----------|----------|
| 1138010104 | 95 | 103 | 107 | | | | 78.5 | 34 | 61.5 |
| 1138010108 | 112 | 101 | 108 | | | | 79 | 45 | 55 |
| 1138010114 | 101 | 102 | 98 | | | | 62 | 39 | 47.5 |
| 1138010128 | 62 | 99 | 124 | | | | 44 | 35 | 40.5 |
| 1138010126 | 98 | 109 | 95 | | | | 59 | 41.5 | 48.5 |
| 1141213504 | 67 | 101 | 92 | | | | 60 | 44 | 39.5 |
| 1138010213 | 70 | 96 | 92 | | | | 59 | 41.5 | 40.5 |
| 1241253209 | 69 | 102 | 76 | 60 | 38 | 57 | | | |

续表

| StuNum | Math | Chinese | English | Physics | Chemistry | Biology | History | Geography | Politics |
|------------|------|---------|---------|---------|-----------|---------|---------|-----------|----------|
| 1138010223 | 51 | 91 | 52 | | | | 70 | 29.5 | 36.5 |
| 1138020411 | 29 | 94 | 71 | | | | 71 | 43.5 | 52 |
| 1141213520 | 25 | 77 | 33 | | | | 38 | 23 | 29 |
| 1241253201 | 104 | 98 | 91 | 72 | 48 | 54 | | | |
| 1138010215 | 91 | 89 | 63 | | | | 71.5 | 27 | 46 |

将数据文件复制到一个文件夹中，例如在“D:\mySAS\”目录下，使用下面命令导入。

```
libname sql v9 "D:\mySAS";
```

1. 选择表中的列——select

使用 select 语句，可以从表中选出想要检索的列。

【例 3-63】 选择 student 表中所有的列。设程序名为 SASTJFX3_63. sas。

```
proc sql outobs =10;  
title 'Information of Student';  
select *  
from sql.student;  
quit;
```

select 后面接 * 表示选择表中所有列。
outobs = 10 是 SAS 中的选项，因为结果较多，限制输出前 10 个结果。
from 后面指定要查询的表名，使用同引用 SAS 数据集一样。
运行结果：

| Information of Student | | | | | |
|------------------------|---------|-----|--------|----|----|
| StuNum | Subject | Sex | School | PE | EG |
| 1138010104 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1138010108 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1138010114 | 文科 | 女 | 北京八中 | 团员 | 汉族 |
| 1138010128 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1138010126 | 文科 | 男 | 北京八中 | 群众 | 汉族 |
| 1141213504 | 文科 | 女 | 北京六中 | 团员 | 回族 |
| 1138010213 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1241253209 | 理科 | 男 | 北京六中 | 团员 | 汉族 |
| 1138010223 | 文科 | 女 | 北京八中 | 团员 | 汉族 |
| 1138020411 | 文科 | 男 | 北京八中 | 团员 | 汉族 |

【例 3-64】 选择特定的列。设程序名为 SASTJFX3_64. sas。

```
proc sql outobs =10;  
title 'Student Number & Subject';  
select StuNum, Subject  
from sql.student;  
quit;
```

select 后面可以写上希望查询的列名称，有多个列时用逗号隔开。
运行结果：

| Student Number & Subject | |
|--------------------------|---------|
| StuNum | Subject |
| 1138010104 | 文科 |
| 1138010108 | 文科 |
| 1138010114 | 文科 |
| 1138010128 | 文科 |
| 1138010126 | 文科 |
| 1141213504 | 文科 |
| 1138010213 | 文科 |
| 1241253209 | 理科 |
| 1138010223 | 文科 |
| 1138020411 | 文科 |

【例 3-65】 消除重复结果。设程序名为 SASTJFX3_65. sas。

```
proc sql;
title 'Information of Schools';
select distinct school
from sql.student;
quit;
```

在 select 后面接 distinct 可以将结果中一样的内容合并在一起。例如希望查找表中同学都分布在哪些学校，而又不希望输出结果中出现重复的名称，可以使用这个选项。

运行结果：

| Information of Schools |
|------------------------|
| School |
| 北京八中 |
| 北京二中 |
| 北京九中 |
| 北京六中 |
| 北京七中 |
| 北京三中 |
| 北京四中 |

2. 创建新的列

仅仅输出表中的各列常常并不能满足要求，有时希望添加一些文字说明列，有时希望将一些列经过简单计算的结果输出。这些问题将会在下方的例子中得到解答。

【例 3-66】 添加新的文字列。设程序名为 SASTJFX3_66. sas。

```
proc sql outobs =10;
select 'The Student',StuNum,'is in school',school
from sql.student;
quit;
```

使用单引号，将要添加的文字引起来，写在 select 后面，同选择多列时的情况一样，需要用逗号隔开。

运行结果：

| | StuNum | | School |
|-------------|------------|--------------|--------|
| The Student | 1138010104 | is in school | 北京八中 |
| The Student | 1138010108 | is in school | 北京八中 |
| The Student | 1138010114 | is in school | 北京八中 |
| The Student | 1138010128 | is in school | 北京八中 |
| The Student | 1138010126 | is in school | 北京八中 |
| The Student | 1141213504 | is in school | 北京六中 |
| The Student | 1138010213 | is in school | 北京八中 |
| The Student | 1241253209 | is in school | 北京六中 |
| The Student | 1138010223 | is in school | 北京八中 |
| The Student | 1138020411 | is in school | 北京八中 |

【例 3-67】 使用列中的信息进行简单计算。设程序名为 SASTJFX3_67. sas。

```
proc sql outobs=10;
select StuNum, Math + Chinese + English as totalScores, (calculated total-
Scores)/3 as averageScores format 5.1
from sql.scores;
quit;
```

如果希望对表中的列进行计算，则直接引用其名即可。默认情况下，新计算出的列是没有名称的，可通过“as”选项定义。如果希望引用计算出来的列，而非原表中的列，需要在引用其名称前加“calculated”。

运行结果：

| Information of Schools | | |
|------------------------|-------------|----------------|
| StuNum | totalScores | average Scores |
| 1138010104 | 305 | 101.7 |
| 1138010108 | 321 | 107.0 |
| 1138010114 | 301 | 100.3 |
| 1138010128 | 285 | 95.0 |
| 1138010126 | 302 | 100.7 |
| 1141213504 | 260 | 86.7 |
| 1138010213 | 258 | 86.0 |
| 1241253209 | 247 | 82.3 |
| 1138010223 | 194 | 64.7 |
| 1138020411 | 194 | 64.7 |

【例 3-68】 分情况产生新列值——case 语句。设程序名为 SASTJFX3_68. sas。

```
proc sql outobs=10;
select StuNum, Math,
       case
         when Math > 135 then 'A'
         when 135 >= Math > 120 then 'B'
         when 120 >= Math > 105 then 'C'
         when 105 >= Math > 90 then 'D'
         when 90 >= Math > 75 then 'E'
```

```

        when 75 >= Math then 'F'
      end as level
    from sql.scores;
  select StuNum, School,
         case School
           when '北京八中' then 'Beijing 8th High School'
           when '北京三中' then 'Beijing 3rd High School'
           else 'none'
         end as numSch
    from sql.student;
quit;

```

例子中首先声明 case，然后对于后面的多种情况分别用“when... then...”语句来描述。“when”后面接情况，“then”后面接要输出的值，这里的值可以是字符串，也可以是数值，规则同添加一个新列一样。

另外，如果情况只涉及等值比较，可以采用本例中第二个查询语句的形式，case 后面为要讨论的变量，这里只能是等值比较，如果需要用不等号，那么只能采用第一种 case 格式。

运行结果：

| StuNum | Math level |
|------------|------------|
| 1138010104 | 95 D |
| 1138010108 | 112 C |
| 1138010114 | 101 D |
| 1138010128 | 62 F |
| 1138010126 | 98 D |
| 1141213504 | 67 F |
| 1138010213 | 70 F |
| 1241253209 | 69 F |
| 1138010223 | 51 F |
| 1138020411 | 29 F |

| StuNum | School | numSch |
|------------|--------|-------------------------|
| 1138010104 | 北京八中 | Beijing 8th High School |
| 1138010108 | 北京八中 | Beijing 8th High School |
| 1138010114 | 北京八中 | Beijing 8th High School |
| 1138010128 | 北京八中 | Beijing 8th High School |
| 1138010126 | 北京八中 | Beijing 8th High School |
| 1141213504 | 北京六中 | none |
| 1138010213 | 北京八中 | Beijing 8th High School |
| 1241253209 | 北京六中 | none |
| 1138010223 | 北京八中 | Beijing 8th High School |
| 1138020411 | 北京八中 | Beijing 8th High School |

【例 3-69】 处理缺失值。设程序名为 SASTJFX3_69.sas。

```

proc sql outobs=10;
select StuNum, Geography, coalesce(Geography,0)
from sql.scores;

```

```

select StuNum, Geography,
       case
         when Geography is missing then 0
         else Geography
       end
from sql.scores;
quit;

```

coalesce 选项可以帮助用户将缺失值替换成希望设定的值。这里需要用户设定两个值：第一个值为希望输出的列，第二个值为前面列在出现缺失时的替代值。

对于缺失的情况，还可以采用 case 语句，如第二个查询语句所示，两个输出结果完全一样。

运行结果：

| StuNum | Geography | |
|------------|-----------|------|
| 1138010104 | 34 | 34 |
| 1138010108 | 45 | 45 |
| 1138010114 | 39 | 39 |
| 1138010128 | 35 | 35 |
| 1138010126 | 41.5 | 41.5 |
| 1141213504 | 44 | 44 |
| 1138010213 | 41.5 | 41.5 |
| 1241253209 | . | 0 |
| 1138010223 | 29.5 | 29.5 |
| 1138020411 | 43.5 | 43.5 |

【例 3-70】 控制输出列名称。设程序名为 SASTJFX3_70. sas。

```

proc sql outobs=10;
select StuNum, StuNum label='#', StuNum label='StudentNumber'
from sql.scores;
quit;

```

当不需要输出原来的列名称或希望更改其名称时，可以使用 label 选项，希望输出空值时可设定 label 值为"#", SAS 不会输出以特殊字符开头的名称。

运行结果：

| StuNum | Student Number | |
|------------|-------------------|------------|
| 1138010104 | 1138010104 | 1138010104 |
| 1138010108 | 1138010108 | 1138010108 |
| 1138010114 | 1138010114 | 1138010114 |
| 1138010128 | 1138010128 | 1138010128 |
| 1138010126 | 1138010126 | 1138010126 |
| 1141213504 | 1141213504 | 1141213504 |
| 1138010213 | 1138010213 | 1138010213 |
| 1241253209 | 1241253209 | 1241253209 |
| 1138010223 | 1138010223 | 1138010223 |
| 1138020411 | 1138020411 | 1138020411 |

3. 数据排序——order

使用“order by”选项可以对查询结果排序。无论是原表中的列，还是生成的新列，都可以进行排序。默认是正序排列，通过设定“desc”可以逆序排列。

【例 3-71】 对多列排序。设程序名为 SASTJFX3_71.sas。

```
proc sql outobs =10;
select StuNum, Math, Chinese, English
from sql.scores
order by Math, Chinese;
quit;
```

如果只设定一个变量，则 SAS 只会依据设定的列排序。如果设定了多个，SAS 则先按第一个排，对于第一个设定变量一样的行，按第二个设定变量进行排序，依此类推。例如，例子中先按“Math”排序后，有 3 个人的数学成绩是一样的，那么就根据第二个设定变量“Chinese”排序。

另外需要注意的是，如果排序列有缺失值，则排在最前。

运行结果：

| StuNum | Math | Chinese | English |
|------------|------|---------|---------|
| 1241223241 | 10 | 46 | 24 |
| 1138020609 | 10 | 87 | 30 |
| 1141213522 | 10 | 95 | 36 |
| 1241223228 | 14 | 76 | 42 |
| 1138020606 | 14 | 92 | 64 |
| 1241223239 | 15 | 49 | 34 |
| 1138020804 | 15 | 82 | 59 |
| 1138020712 | 15 | 96 | 34 |
| 1241223234 | 16 | 82 | 23 |
| 1141213525 | 17 | 70 | 36 |

【例 3-72】 对三门主课(语数外)平均成绩按由大到小排序。设程序名为 SASTJFX3_72.sas。

```
proc sql outobs =10;
select StuNum, Math + Chinese + English as totalScores, (calculated totalScores)/3 as averageScores format 5.1
from sql.scores
order by averageScores desc, StuNum;
quit;
```

这里使用了计算生成的新列进行排序，在“order by”后使用计算得出的列，不需要声明“calculated”，而且仅在此处不需要，其他语句后面引用计算得出的结果都需要声明“calculated”。在这个例子中，对平均成绩按从大到小排序，使用“desc”选项，跟在“averageScores”之后。

运行结果：

| StuNum | totalScores | average Scores |
|------------|-------------|-------------------|
| 1138010108 | 321 | 107.0 |
| 1138010106 | 310 | 103.3 |
| 1138010102 | 307 | 102.3 |

| | | |
|------------|-----|-------|
| 1138010103 | 307 | 102.3 |
| 1138010104 | 305 | 101.7 |
| 1138010126 | 302 | 100.7 |
| 1138010114 | 301 | 100.3 |
| 2138020118 | 297 | 99.0 |
| 1241253201 | 293 | 97.7 |
| 1138010128 | 285 | 95.0 |

4. 检索满足特定要求的数据——where

有时并不需要输出表中所有数据，而只希望检索出满足要求的一部分。
先看一个例子以了解 where 的基本用法。

【例 3-73】 检索所有文科生。设程序名为 SASTJFX3_73.sas。

```
proc sql outobs =10;  
select StuNum, Subject  
from sql.student  
where Subject = "文科";  
quit;
```

运行结果：

| StuNum | Subject |
|------------|---------|
| 1138010104 | 文科 |
| 1138010108 | 文科 |
| 1138010114 | 文科 |
| 1138010128 | 文科 |
| 1138010126 | 文科 |
| 1141213504 | 文科 |
| 1138010213 | 文科 |
| 1138010223 | 文科 |
| 1138020411 | 文科 |
| 1141213520 | 文科 |

前面在介绍 case 语句时，用到了一些比较运算符，这里系统地介绍 SAS 中 SQL 过程涉及的运算符，参见表 3-22。

表 3-22 SAS 中 SQL 过程涉及的运算符

| 符 号 | 等价缩写 | 定 义 | 举 例 | 备 注 |
|-------------------------------|------|------|-----------------------------------|-----|
| 比较运算符 | | | | |
| = | EQ | 相等 | where Name = 'Asia' ; | |
| ^ = or ~ = | NE | 不等 | where Name ne 'Africa' ; | |
| or \sqsupset = or \langle | | | | |
| > | GT | 大于 | where Area > 10000 ; | |
| < | LT | 小于 | where Depth < 5000 ; | |
| >= | GE | 大于等于 | where Statehood >= '01jan1860'd ; | |
| < = | LE | 小于等于 | where Population <= 5000000 ; | |

续表

| 符 号 | 等价 缩写 | 定 义 | 举 例 | 备 注 |
|--------------------------|----------|--|---|--|
| 逻辑运算符 | | | | |
| & | AND | 逻辑与 | Continent = ' Asia ' and Population >5000000 | |
| ! or or | OR | 逻辑或 | Population < 1000000 or Population >5000000 | |
| ^ or ~ or ¬ | NOT | 逻辑非 | Continent not ' Africa ' | |
| 条件运算符 | | | | |
| ANY | | 在子查找中至少存在一个满足条件的结果 | where Population > any (select Popula- tion from sql. countries) | 查找比表 countries 中任意最小的 Population 值大的元组 |
| ALL | | 在子查找中的所有结果均满足条件要求 | where Population > all(select Population from sql. countries) | 查找比表 countries 中任意最大的 Population 值大的元组 |
| BETWEEN- AND | | 属性在一个区间内的元组 | where Population between 1000000 and 5000000 | 查找 Population 介于 100 万~500 万之间的元组 |
| CONTAINS | | 查找是否包含特定的字符串 | where Continent contains ' America ' ; | 查找 Continent 中包含 America 的元组 |
| EXISTS | | 在子查找中查找是否存在，不返回数据，只产生逻辑真值或逻辑假值 | where exists(select * from sql. oilprod) ; | |
| IN | | 查找属性值属于列出的指定集合的元组 | where Name in(' Africa ' , ' Asia ') ; | 查找 Name 中有“ Africa ”或“ Asia ”的元组 |
| IS NULL or IS MISSING | | 判断是否是缺失值 | where Population is missing ; | 查找 Population 缺失的元组 |
| LIKE | | 字符串匹配，可包含通配符，“%”代表任意字符串，“_”代表一个字符。如表示一个中文字符，应使用两个“_” | where Continent like ' A% ' ; | 查找 continent 中以 A 开头的元组 |
| = * | | 匹配发音与指定字符串相近的字符串 | where Name = * ' Tiland ' | 查找发音同“ Tiland ”的字符串 |
| 字符串比较运算符 | | | | |
| EQT | | 与指定字符串相等 | where Name eqt ' Aust ' ; | |
| GTT | | 比指定字符串大 | where Name gtt ' Bah ' ; | |
| LTT | | 比指定字符串小 | where Name ltt ' An ' ; | |
| GET | | 大于或等于字符串 | where Country get ' United A ' ; | |
| LET | | 小于或等于字符串 | where Lastname let ' Smith ' ; | |
| NET | | 不等于字符串 | where Style net ' TWO ' ; | |

【例 3-74】 复合条件查询——查询语文成绩和数学成绩同时大于 90 分的学生。设程序名为 SASTJFX3_74. sas。

```
proc sql;
select StuNum,chinese,math
from sql.scores
where chinese>90 & math>90;
quit;
```

当需要设定多个条件时，需要同时使用逻辑运算符和比较运算符。这里没有输出这些学生的语文成绩和数学成绩，没有影响到正常查询。

运行结果：

| StuNum | Chinese | Math |
|------------|---------|------|
| 1138010104 | 103 | 95 |
| 1138010108 | 101 | 112 |
| 1138010114 | 102 | 101 |
| 1138010126 | 109 | 98 |
| 1241253201 | 98 | 104 |
| 1141213505 | 98 | 99 |
| 1138010106 | 104 | 105 |
| 1138010102 | 107 | 98 |

【例 3-75】 使用 IN 条件操作符——查询表中北京三中和北京四中的学生。设程序名为 SASTJFX3_75.sas。

```
proc sql;
select StuNum
from sql.student
where school in ("北京三中", "北京四中");
quit;
```

“IN”后面紧跟的括号中列出要查找的内容。括号内也可以为子查询，可用于多表查询，这会在后面的章节中用到。与 IN 相对的是“NOT IN”，表示不在指定查找内容的范围内。

运行结果：

| StuNum |
|------------|
| 1241223210 |
| 1138010313 |
| 1138020517 |

【例 3-76】 空值可能引起的问题。设程序名为 SASTJFX3_76.sas。

```
proc sql;
select StuNum, Physics
from sql.scores
where Physics < 50
order by Physics;
select StuNum, Physics
from sql.scores
where Physics < 50 and Physics is not missing
order by Physics;
quit;
```

当使用比较运算符小于时，如果有空值，如例子中所示，都会包含在查找结果中。有时，我们并不需要这些结果。例子中的第二个查询使用了“is not missing”，这样就可以把空值排除在外。

运行结果如下：

第一个查询(只包含前面 10 个结果)：

| StuNum | Physics |
|------------|---------|
| 1138010313 | . |
| 1138020804 | . |
| 1141213505 | . |
| 1138020718 | . |
| 1141213525 | . |
| 1138010126 | . |
| 1138020424 | . |
| 1138020726 | . |
| 1138010205 | . |
| 1141213514 | . |

第二个查询：

| StuNum | Physics |
|------------|---------|
| 1241223236 | 8.5 |
| 1241223234 | 10 |
| 1241253222 | 12 |
| 1241223227 | 14 |
| 1241223239 | 15 |
| 1241223231 | 17 |
| 1241223225 | 18.5 |
| 1241223210 | 19.5 |
| 1241223232 | 20.5 |
| 1241223235 | 21 |
| 1241223241 | 22 |
| 1241223211 | 26 |
| 1241223237 | 26 |
| 1241223242 | 28 |
| 1241233205 | 30 |
| 1241233212 | 35.5 |
| 1241223228 | 36 |
| 1241223223 | 36 |
| 1241253216 | 41 |
| 1241253220 | 45 |
| 1241253219 | 46.5 |
| 1241233218 | 48 |

5. 聚集函数

SQL 过程有丰富的聚集函数(见表 3-23)，进一步方便用户，增强检索功能。

表 3-23 SQL 聚集函数

| 函 数 名 | 功 能 | 函 数 名 | 功 能 |
|----------------|-----------------|--------|---------------------------|
| AVG, MEAN | 均值 | RANGE | 值的范围 |
| COUNT, FREQ, N | 统计个数，不计算空值的项 | STD | 标准差 |
| CSS | 修正的平方和 | STDERR | 平均数标准误差 |
| CV | 变异系数 | SUM | 总和 |
| MAX | 最大值 | SUMWGT | 各变量权重的总和，SQL 中每行数据的权重均为 1 |
| MIN | 最小值 | T | 假设检验与 0 比较的 T 值 |
| NMISS | 空值的数量 | USS | 未修正的平方和 |
| PRT | 假设检验与 0 比较的 P 值 | VAR | 方差 |

【例 3-77】 查询学生总人数和表中涉及的学校数量。设程序名为 SASTJFX3_77.sas。

```
proc sql;
select count(*) label = "Num of Students", count(distinct school) label
      = "Num of Schools"
from sql.student;
quit;
```

运行结果：

| Num of Students | Num of Schools |
|--------------------|-------------------|
| 100 | 7 |

【例 3-78】 查询所有人的物理成绩平均分。设程序名为 SASTJFX3_78.sas。

```
/* 计算物理平均成绩* /
proc sql;
select avg(Physics) as AvgPhysics
from sql.scores
quit;
proc sql;
/* 特殊情况, 考虑空值* /
select StuNum, case
      when Physics is missing then 0
      else Physics
      end as NewPhysics,
      avg(calculated NewPhysics) as AvgPhysics
from sql.scores
quit;
```

因为表里包括文科生和理科生的成绩，文科生的物理成绩均为空值，计算是需要排除空值的。SQL 已经考虑到这一点，对于空值是不会计入的。实际上，在聚集函数遇到空值时，除了 count 以外，都跳过空值而只处理非空值。

假如在某种特殊情况下，录入数据时将 0 都记为空值，在计算平均数等时，会跳过空值行，不计在总数内。如果希望考虑这些空值，对于本例，可以按如下方式查询。

运行结果如下。

第一个查询结果：

| AvgPhysics |
|------------|
| 34.19643 |

第二个查询结果：

| StuNum | NewPhysics | AvgPhysics |
|------------|------------|------------|
| 1241223211 | 26 | 9.575 |
| 1241223235 | 21 | 9.575 |
| 1138010119 | 0 | 9.575 |
| 1241253216 | 41 | 9.575 |
| 1141213510 | 0 | 9.575 |
| 1241223231 | 17 | 9.575 |

6. 为数据分组——Group By

这个语句在 SAS 过程中经常用到，在这里，其作用也是一样的。通过“Group By”语句，可以将数据按某列不同的值进行分组，这里还可以使用聚集函数，分别计算每组数据。

【例 3-79】 统计学生的政治面貌。设程序名为 SASTJFX3_79.sas。

```
proc sql;
select PE, count(*)
from sql.student
group by PE;
quit;
```

运行结果：

| PE | |
|----|----|
| 群众 | 15 |
| 团员 | 85 |

【例 3-80】 统计不同学校的学生的政治面貌。设程序名为 SASTJFX3_80.sas。

```
proc sql;
select School, PE, count(*)
from sql.student
group by School, PE;
quit;
```

这里涉及多个变量的分组，类似对单个分组，在“Group By”后加要分组的两个变量。

运行结果：

| School | PE | |
|--------|----|----|
| 北京八中 | 群众 | 8 |
| 北京八中 | 团员 | 44 |
| 北京二中 | 群众 | 1 |
| 北京二中 | 团员 | 1 |
| 北京九中 | 团员 | 1 |
| 北京六中 | 群众 | 6 |
| 北京六中 | 团员 | 35 |
| 北京七中 | 团员 | 1 |
| 北京三中 | 团员 | 2 |
| 北京四中 | 团员 | 1 |

7. 过滤分组查询结果——HAVING

“Having”语句与“Group by”一起使用，其主要作用是过滤分组好的数据，挑选出所需要的分组。

【例 3-81】 统计北京八中和北京六中的学生政治面貌。设程序名为 SASTJFX3_81.sas。

```
proc sql;
select School, PE, count(*)
from sql.student
group by School, PE
having School in ("北京八中", "北京六中");
quit;
```

运行结果：

| School | PE | |
|--------|----|----|
| 北京八中 | 群众 | 8 |
| 北京八中 | 团员 | 44 |
| 北京六中 | 群众 | 6 |
| 北京六中 | 团员 | 35 |

可以看到，“Having”同“Where”十分类似。下面对这二者进行一个比较，见表 3-24。

表 3-24 HAVING 与 WHERE 语句的比较

| HAVING | WHERE |
|---|----------------------------|
| 对组设定选择条件，用来包含或删除指定的组 | 对数据元组设定选择条件，用来包含或删除指定的数据元组 |
| 必须放在“Group By”语句之后 | 必须放在“Group By”语句之前 |
| 受“Group By”语句影响，若没有“Group By”语句，则与“Where”相同 | 不受“Group By”语句影响 |
| 在“Group By”和聚集函数之后运行 | 在“Group By”和聚集函数之前运行 |

【例 3-82】 统计北京八中和北京六中汉族学生的政治面貌。设程序名为 SASTJFX3_82.sas。

```
proc sql;
select School, PE, count(*)
from sql.student
where EG="汉族"
group by School, PE
having School in("北京八中","北京六中");
quit;
```

运行结果：

| School | PE | |
|--------|----|----|
| 北京八中 | 群众 | 7 |
| 北京八中 | 团员 | 42 |
| 北京六中 | 群众 | 6 |
| 北京六中 | 团员 | 30 |

8. 多表连接查询

当要输出的信息涉及多个表时，需要连接操作实现。所谓的连接操作，就是将多个表的信息，通过一定规则条件有选择地组合在一起。

在了解 SQL 的连接操作前，需要先了解笛卡儿积的定义。

定义 1 域(Domain)：一组具有相同数据类型的值的集合。

例如，自然数、整数、实数等，在 SAS 表中，每一列数据一般都具有相同的属性，也可看成一个域。

定义 2 笛卡儿积(Cartesian Product)：给定一组域 D1, D2, ..., Dn, 这些域可以是相同的域。D1, D2, ..., Dn 的笛卡儿积为

$$D1 * D2 * \dots * Dn = \{ (d1, d2, \dots, dn) | di \in Di, i = 1, 2, \dots, n \}$$

这相当于 n 维域的所有元素的所有组合。更直观的可以看下面这个例子。

【例 3-83】 用 SQL 求笛卡儿积。设程序名为 SASTJFX3_83.sas。

先给出三个域：

```
D1 = {A1,A2}
D2 = {B1,B2,B3}
D3 = {C1,C2,C3}
```

求 D1, D2, D3 的笛卡儿积 SAS 程序：

```
data d1;
input x1 $ @@ ;
cards;
A1 A2
;
data d2;
input x2 $ @@ ;
cards;
B1 B2 B3
;
data d3;
input x3 $ @@ ;
cards;
C1 C2 C3
;
proc sql;
title "Cartesian Product";
select *
from d1,d2,d3;
quit;
```

运行结果：

| Cartesian Product | | |
|-------------------|----|----|
| x1 | x2 | x3 |
| A1 | B1 | C1 |
| A1 | B2 | C1 |
| A1 | B3 | C1 |
| A2 | B1 | C1 |
| A2 | B2 | C1 |
| A2 | B3 | C1 |
| A1 | B1 | C2 |
| A1 | B2 | C2 |
| A1 | B3 | C2 |
| A2 | B1 | C2 |
| A2 | B2 | C2 |
| A2 | B3 | C2 |
| A1 | B1 | C3 |
| A1 | B2 | C3 |
| A1 | B3 | C3 |
| A2 | B1 | C3 |
| A2 | B2 | C3 |
| A2 | B3 | C3 |

因此可知，D1、D2、D3 的笛卡儿积为：

{ (A1,B1,C1), (A1,B2,C1), (A1,B3,C1), (A2,B1,C1), (A2,B2,C1), (A2,B3,C1), (A1,B1,C2), (A1,B2,C2), (A1,B3,C2), (A2,B1,C2), (A2,B2,C2), (A2,B3,C2), (A1,B1,C3), (A1,B2,C3), (A1,B3,C3), (A2,B1,C3), (A2,B2,C3), (A2,B3,C3) }

由于连接操作比较复杂，为方便理解，引入四个数据集。本节后面的例子都会用到这几个，不再特别列出。

| 数据集 one | | |
|---------|----|----|
| A1 | B1 | C1 |
| A1 | B2 | C2 |
| A2 | B2 | C1 |
| A3 | B1 | D1 |

| 数据集 two | | |
|---------|----|----|
| A1 | B2 | C2 |
| A1 | B3 | C2 |
| A2 | B2 | C1 |
| A2 | B2 | D1 |

| 数据集 three | | |
|-----------|---|---|
| 1 | 2 | 1 |
| 2 | 3 | 2 |
| 3 | 3 | 3 |
| 4 | 4 | 4 |

| 数据集 four | | |
|----------|---|---|
| 1 | 2 | 1 |
| 2 | 2 | 2 |
| 3 | 3 | 3 |
| 4 | 1 | 4 |

【例 3-84】 表与表之间的笛卡儿积。设程序名为 SASTJFX3_84.sas。

```
proc sql;
title "Cartesian Product ";
select * from one, two;
select * from one cross join two;
quit;
```

在前面域的定义中，其中每个元素并不一定代表都是一个单一的值。在表与表之间的笛卡儿积中，每个表的一行整体作为象征意义上的一个值。两个查询语句结果完全一样，都是求笛卡儿积。

运行结果：

| Cartesian Product | | | | | |
|-------------------|----|----|----|----|----|
| X | Y | Z | M | N | P |
| A1 | B1 | C1 | A1 | B2 | C2 |
| A1 | B1 | C1 | A1 | B3 | C2 |
| A1 | B1 | C1 | A2 | B2 | C1 |
| A1 | B1 | C1 | A2 | B2 | D1 |
| A1 | B2 | C2 | A1 | B2 | C2 |
| A1 | B2 | C2 | A1 | B3 | C2 |
| A1 | B2 | C2 | A2 | B2 | C1 |
| A1 | B2 | C2 | A2 | B2 | D1 |
| A2 | B2 | C1 | A1 | B2 | C2 |
| A2 | B2 | C1 | A1 | B3 | C2 |
| A2 | B2 | C1 | A2 | B2 | C1 |
| A2 | B2 | C1 | A2 | B2 | D1 |
| A3 | B1 | D1 | A1 | B2 | C2 |
| A3 | B1 | D1 | A1 | B3 | C2 |
| A3 | B1 | D1 | A2 | B2 | C1 |
| A3 | B1 | D1 | A2 | B2 | D1 |

【例 3-85】 内连接——笛卡儿积的子集。设程序名为 SASTJFX3_85.sas。

笛卡儿积的结果往往十分多，并非全都需要。通过限定条件，可以选出所需的结果。例如，我们只需要表 one 和表 two 中 X 和 M 相等的结果。

```
proc sql;
  title "Inner Joins";
  select * from one as t1, two as t2
  where t1.X=t2.M;
quit;
```

运行结果：

| Inner Joins | | | | | |
|-------------|----|----|----|----|----|
| X | Y | Z | M | N | P |
| A1 | B1 | C1 | A1 | B2 | C2 |
| A1 | B1 | C1 | A1 | B3 | C2 |
| A1 | B2 | C2 | A1 | B2 | C2 |
| A1 | B2 | C2 | A1 | B3 | C2 |
| A2 | B2 | C1 | A2 | B2 | C1 |
| A2 | B2 | C1 | A2 | B2 | D1 |

这里有两种方式可以实现内连接查询：一种是通过“where”语句，另一种是通过“inner join ... on”语句。两者结果一致。语句如下：

```
select * from one t1 inner join two t2 on t1.X=t2.M;
```

在前面的小节中，提到给变量重命名。这里，可以通过如例子中所示的方式给表 one 和表 two 重命名为 t1 和 t2。完整的重命名格式需要在原表名和重命名之间加“as”，这里的“as”是可以省略的。因此，第一个查询中使用了“as”，而第二个查询没有使用，但结果一样。

我们看到，在结果中，X 和 M 的值一直是一样的。如果只想输出其中一个，可以将查询语句修改为：

```
select t1.X,t1.Y,t1.Z,t2.N,t2.P from one t1 innner join two t2 on t1.X=t2.M;
```

从这几个例子中可以看到，如果在多表查询中希望选择或对某一个表项的值引用，需要使用“表.属性”的格式。如果所连接的表的变量没有重名的，那么前面的表名也可以省略，例如上面的查询语句可简化为：

```
select X,Y,Z,N,P from one innner join two on X=M;
```

运行结果：

| Inner Joins | | | | |
|-------------|----|----|----|----|
| X | Y | Z | N | P |
| A1 | B1 | C1 | B2 | C2 |
| A1 | B1 | C1 | B3 | C2 |
| A1 | B2 | C2 | B2 | C2 |
| A1 | B2 | C2 | B3 | C2 |
| A2 | B2 | C1 | B2 | C1 |
| A2 | B2 | C1 | B2 | D1 |

使用条件操作符，可以不单单限定一个条件，例如如果再加入 Y 和 N 相等的条件，则查询语句和结果为：

```
select * from one inner join two on X=M and Y=N;
```

运行结果：

| Inner Joins | | | | | |
|-------------|----|----|----|----|----|
| X | Y | Z | M | N | P |
| A1 | B2 | C2 | A1 | B2 | C2 |
| A2 | B2 | C1 | A2 | B2 | C1 |
| A2 | B2 | C1 | A2 | B2 | D1 |

【例 3-86】 左外连接和右外连接。设程序名为 SASTJFX3_86.sas。

在上面的例子中，内连接会扔掉所有不符合条件限制的笛卡儿积结果。有时，我们仍需要保留那些没有匹配到的项，这时需要用到外连接。

当只涉及两个表时，若希望保留第一个表中的所有项，则使用左外连接。

```
proc sql;
title "Left Outer Joins";
select *
from one t1 left join two t2 on t1.X=t2.M and t1.Y=t2.N;
quit;
```

运行结果：

| Left Outer Joins | | | | | |
|------------------|----|----|----|----|----|
| X | Y | Z | M | N | P |
| A1 | B1 | C1 | | | |
| A1 | B2 | C2 | A1 | B2 | C2 |
| A2 | B2 | C1 | A2 | B2 | C1 |
| A2 | B2 | C1 | A2 | B2 | D1 |
| A3 | B1 | D1 | | | |

根据结果可以看出，左外连接除了将所有内连接的结果输出，还将表 one 中其他数据输出，相应的 M、N、P 值用空值表示。

右外连接同理。

```
proc sql;
title "Right Outer Joins";
select *
from one t1 right join two t2 on t1.X=t2.M and t1.Y=t2.N;
quit;
```

运行结果：

| Right Outer Joins | | | | | |
|-------------------|----|----|----|----|----|
| X | Y | Z | M | N | P |
| A1 | B2 | C2 | A1 | B2 | C2 |
| | | | A1 | B3 | C2 |
| A2 | B2 | C1 | A2 | B2 | C1 |
| A2 | B2 | C1 | A2 | B2 | D1 |

【例 3-87】 全外连接。设程序名为 SASTJFX3_87.sas。

全外连接是左外连接和右外连接的组合。

```
proc sql;
title "Full Outer Joins";
select *
from one t1 full join two t2 on t1.X=t2.M and t1.Y=t2.N;
quit;
```

运行结果：

| Full Outer Joins | | | | | |
|------------------|----|----|----|----|----|
| X | Y | Z | M | N | P |
| A1 | B1 | C1 | | | |
| A1 | B2 | C2 | A1 | B2 | C2 |
| | | | A1 | B3 | C2 |
| A2 | B2 | C1 | A2 | B2 | C1 |
| A2 | B2 | C1 | A2 | B2 | D1 |
| A3 | B1 | D1 | | | |

【例 3-88】 Union 连接。设程序名为 SASTJFX3_88.sas。

```
proc sql;
title "Union Join";
select *
from three union join four;
quit;
```

使用 Union 连接时，两表不会进行比较，而是直接将两个表合并在一起，总行数为两表行数之和。双方的行之间没有交集。

运行结果：

| Union Join | | | | | |
|------------|---|---|---|---|---|
| X | Y | Z | X | Y | W |
| . | . | . | 1 | 2 | 1 |
| . | . | . | 2 | 2 | 2 |
| . | . | . | 3 | 3 | 3 |
| . | . | . | 4 | 1 | 4 |
| 1 | 2 | 1 | . | . | . |
| 2 | 3 | 2 | . | . | . |
| 3 | 3 | 3 | . | . | . |
| 4 | 4 | 4 | . | . | . |

【例 3-89】 自然连接。设程序名为 SASTJFX3_89.sas。

自然连接是一个方便查询的便捷功能。程序会自动在查询的几个表中根据列名称和数据类型找到一样的，进行等值匹配。

```
proc sql;
title "Natural Join";
select *
from three natural join four;
quit;
```

例子中，表 three 和表 four 有共同的变量 X 和 Y，自然连接会自动等值匹配两个表中的 X 和 Y，相当于使用了“where three. X = four. X and three. Y = four. Y”的条件。

运行结果：

| Natural Join | | | |
|--------------|---|---|---|
| X | Y | W | Z |
| 1 | 2 | 1 | 1 |
| 3 | 3 | 3 | 3 |

【例 3-90】 自身连接。设程序名为 SASTJFX3_90. sas。

有时会需要表与表自身进行连接，这时要用到同一表的重命名。

```
proc sql;
title "Self Join";
select a.X,a.Y,a.Z,' |',b.Y,b.X,b.Z
from three a,three b
where a.X=b.Y;
quit;
```

运行结果：

| Self Join | | | | | | |
|-----------|---|---|--|---|---|---|
| X | Y | Z | | Y | X | Z |
| 2 | 3 | 2 | | 2 | 1 | 1 |
| 3 | 3 | 3 | | 3 | 2 | 2 |
| 3 | 3 | 3 | | 3 | 3 | 3 |
| 4 | 4 | 4 | | 4 | 4 | 4 |

9. 嵌套查询

在 SQL 中，一个 SELECT - FROM - WHERE 语句成为一个查询块。将一个查询块嵌套在另一个查询块的 WHERE 子句或 HAVING 短语的条件中的查询称为嵌套查询。

嵌套查询中最常用到语句“in”、“Any”、“All”、“Exists”以及比较操作符。

【例 3-91】 求比北京六中学生语文成绩最高分高的学生以及其所在学校。设程序名为 SASTJFX3_91. sas。

```
proc sql;
select s1.StuNum,s1.school,s2.chinese
from sql.student s1, sql.scores s2
where s1.StuNum=s2.StuNum
and s2.chinese > (
select Max(s4.chinese)
from sql.student s3, sql.scores s4
where s3.StuNum=s4.StuNum and s3.school = '北京六中');
quit;
```

运行结果：

| StuNum | School | Chinese |
|------------|--------|---------|
| 1138010126 | 北京八中 | 109 |
| 1138010205 | 北京八中 | 114 |
| 1138020601 | 北京八中 | 110 |

| | | |
|------------|------|-----|
| 1138010103 | 北京八中 | 122 |
| 1138010129 | 北京八中 | 112 |

SQL 在进行嵌套查询时，会从最内层的查询开始。子查询先生成一个新的表，将学生表和成绩表根据学号内连接，并挑出所有北京六中的学生。再使用聚集函数 MAX 求得最高的语文成绩。这样，子查询会返回一个值，用于父查询条件限制。

父查询中，同样通过内连接将学生表和成绩表连接在一起，然后通过比较筛选出语文成绩比子查询返回值大的学生。

这段查询还可以使用“ALL”语句实现相同的功能，可以不使用聚集函数“MAX”。但是，用聚集函数实现子查询通常比直接用 Any 或 All 查询的效率高，建议使用第一种做法。

```
select s1.StuNum,s1.school,s2.chinese
from sql.student s1, sql.scores s2
where s1.StuNum=s2.StuNum
      and s2.chinese > all (
      select s4.chinese
      from sql.student s3, sql.scores s4
      where s3.StuNum=s4.StuNum and s3.school = '北京六中');
```

【例 3-92】 查询英语成绩大于 100 的学生信息。设程序名为 SASTJFX3_92.sas。

本例的子查询的返回值不再是单一的一个值，而是一个集合。

```
proc sql;
select *
from sql.student s1
where s1.StuNum in (
      select s2.StuNum
      from sql.scores s2
      where s2.english > 100);
quit;
```

运行结果：

| StuNum | Subject | Sex | School | PE | EG |
|------------|---------|-----|--------|----|----|
| 1138010104 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1138010108 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1138010128 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1138010222 | 文科 | 女 | 北京八中 | 团员 | 汉族 |
| 2138020118 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1138010106 | 文科 | 女 | 北京八中 | 团员 | 汉族 |
| 1241253208 | 理科 | 女 | 北京六中 | 团员 | 汉族 |
| 1141213502 | 文科 | 男 | 北京六中 | 团员 | 壮族 |
| 1138010103 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1138010102 | 文科 | 男 | 北京八中 | 团员 | 汉族 |

子查询在成绩表中将英语成绩大于 100 的学生挑选出来生成一个新的学号的集合。在父查询中，只需将学号属于子查询返回的集合的学生信息输出即可。

【例 3-93】 使用 EXISTS 的样例。设程序名为 SASTJFX3_93.sas。

一些带 EXISTS 或 NOT EXISTS 的子查询不能被其他形式的子查询替换，但所有带 IN、比较运算符、ANY 和 ALL 的子查询都能被带 EXISTS 的子查询等价替换。

```
proc sql;
select *
from sql.student s1
where exists(
    select *
    from sql.scores s2
    where s2.english>100 and s1.StuNum=s2.StuNum);
quit;
```

运行结果:

| StuNum | Subject | SexSchool | PE | EG | |
|------------|---------|-----------|------|----|----|
| 1138010104 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1138010108 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1138010128 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1138010222 | 文科 | 女 | 北京八中 | 团员 | 汉族 |
| 2138020118 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1138010106 | 文科 | 女 | 北京八中 | 团员 | 汉族 |
| 1241253208 | 理科 | 女 | 北京六中 | 团员 | 汉族 |
| 1141213502 | 文科 | 男 | 北京六中 | 团员 | 壮族 |
| 1138010103 | 文科 | 男 | 北京八中 | 团员 | 汉族 |
| 1138010102 | 文科 | 男 | 北京八中 | 团员 | 汉族 |

10. 查询结果操作符

SQL 提供了将两个或多个查询结果结合的操作符, 共有以下 4 个:

- (1) UNION 求并集。
- (2) INTERSECT 求交集。
- (3) EXCEPT 集合减法, 去掉第一个集合中与第二个集合的交集。
- (4) OUTER UNION 从每个查询结果中产生单独的行。

【例 3-94】 两个数据集的不同连接。设程序名为 SASTJFX3_94.sas。

```
proc sql;
title "oneUnion two";
select * from one
union
select * from two;
title "one Intersect two";
select * from one
intersect
select * from two;
title "one Except two";
select * from one
except
select * from two;
title "one Outer Union two";
select * from one
outer union
select * from two;
quit;
```

运行结果:

| oneUnion two | | | one Intersect two | | |
|--------------|----|----|-------------------|----|----|
| X | Y | Z | X | Y | Z |
| A1 | B1 | C1 | A1 | B2 | C2 |
| A1 | B2 | C2 | A2 | B2 | C1 |
| A1 | B3 | C2 | | | |
| A2 | B2 | C1 | | | |
| A2 | B2 | D1 | | | |
| A3 | B1 | D1 | | | |

| one Except two | | | oneExcept two | | |
|----------------|----|----|---------------|----|----|
| X | Y | Z | X | Y | Z |
| A1 | B1 | C1 | A1 | B1 | C1 |
| A3 | B1 | D1 | A3 | B1 | D1 |

结果解释：

数据集 one 和 two 为例 3-84 之前提到的四个数据集中的前两个。

UNION 结果：虽然两个集合的变量名不同，但查询结果操作符是根据内容选择的，因此，最终的结果为两者的并集，默认使用了第一个集合的变量名。

INTERSECT 结果：两个表共有两个相同行，求交集的结果即这两个相同的行。

EXCEPT 结果：表 one 中的数据去掉上一步求交集的结果，即最终查询结果。

OUTER UNION 结果：不管两个表中的数据是什么样的，每个表的数据行都在结果中生成一个独立的行。

11. 使用 SQL 创建新表

在 SAS 中，创建新表就相当于创建数据集。SQL 为我们提供了一个非常方便的查询数据的工具，而对于数据集的创建和数据导入，还是数据步更方便一些。一般，我们使用 SQL 创建新的表，都是基于其查询结果。

在上面众多的例子中，输出都直接显示在结果窗口中，而我们常常需要将这些结果做进一步处理，可以使用“create”语句来实现。

【例 3-95】 从查询结果创建新表。设程序名为 SASTJFX3_95.sas。

```
proc sql;
create table queryResult as
select *
from sql.student s1
where s1.StuNum in(
select s2.StuNum
from sql.scores s2
where s2.english>100);
quit;
```

沿用例 3-92 的查询语句，将结果直接创建一个新的表。由于没有指定新表的逻辑库，因此，在 SAS 的“work”库中会出现一个新数据集“Queryresult”。

这里 create 后的 table 选项是指创建一个 SAS 数据集。如果写为 view，则创建一个视图。在希望节省空间、保持新生成的表能随原数据同步更新等情况下，选用创建 view。

相对于创建，删除表很简单。

```
proc sql;
drop table tablename;
```

其中，“tablename”为所要删除的表名。

12. 添加新的数据行

添加新的数据有两种方式：一种是直接添加信息，另一种是添加查询结果。

【例 3-96】 使用 set 直接添加信息。设程序名为 SASTJFX3_96.sas。

```
data insdata;
input x y $ z;
cards;
1 A 2
3 B 4
5 C 6
;
proc sql;
insert into insdata
    set x=7,y='D',z=8
    set x=9,y='E',z=10;
title "Inserted dataset";
select * from insdata;
quit;
```

运行结果：

| Inserted dataset | | |
|------------------|---|----|
| x | y | z |
| 1 | A | 2 |
| 3 | B | 4 |
| 5 | C | 6 |
| 7 | D | 8 |
| 9 | E | 10 |

【例 3-97】 使用 values 直接添加信息。设程序名为 SASTJFX3_97.sas。

```
data insdata;
input x y $ z;
cards;
1 A 2
3 B 4
5 C 6
;
proc sql;
insert into insdata
values(7,'D',8)
values(., 'E',10);
title "Inserted dataset";
select * from insdata;
quit;
```

运行结果：

| Inserted dataset | | |
|------------------|---|----|
| x | y | z |
| 1 | A | 2 |
| 3 | B | 4 |
| 5 | C | 6 |
| 7 | D | 8 |
| . | E | 10 |

【例 3-98】 利用查询结果添加信息。设程序名为 SASTJFX3_98.sas。

```
data insdata;
input x y $ z;
cards;
1 A 2
3 B 4
5 C 6
;
data desdata;
input x y $ z;
cards;
10 G 12
9 F 8
6 Q 7
;
proc sql;
insert into desdata
select * from insdata where x>3;
title "Inserted dataset";
select * from desdata;
quit;
```

运行结果:

| Inserted dataset | | |
|------------------|---|----|
| x | y | z |
| 10 | G | 12 |
| 9 | F | 8 |
| 6 | Q | 7 |
| 5 | C | 6 |

13. 更新数据

在 SAS 的 data 步中, 进行整列的操作很方便, 在 SQL 里也可以很方便地实现。

【例 3-99】 修改整列数据。设程序名为 SASTJFX3_99.sas。

在 SAS 的 data 步中, 对满足特定条件的数据进行修改是件十分麻烦的事情, 这时, 就可以借助 SQL 来简单实现。

```
data u1;
input x y $ z;
cards;
1 A 2
3 B 4
5 C 6
;
```

```
proc sql;
update ul
    set z = z* 1.2;
title "Update data";
select * from ul;
quit;
```

运行结果:

| Update data | | |
|-------------|---|-----|
| x | y | z |
| 1 | A | 2.4 |
| 3 | B | 4.8 |
| 5 | C | 7.2 |

利用 SQL 语句, 将 z 的数值都乘以 1.2。

【例 3-100】 修改特定数据。设程序名为 SASTJFX3_100. sas。

```
data u2;
input x y $ z;
cards;
1 Ad 2
1 Ae 1
2 Afg 3
2 Bq 4
2 Bm 9
3 Cz 6
;
proc sql;
update u2
    set z = z* 1.2
    where y like "A%" and x < 2 ;
update u2
    set x = x*
    case
        when y like "A%" then 2
        when y like "B%" then 3
        else 4
    end;
title "Update data";
select * from u2;
quit;
```

运行结果:

| Update data | | |
|-------------|-----|-----|
| x | y | z |
| 2 | Ad | 2.4 |
| 2 | Ae | 1.2 |
| 4 | Afg | 3 |
| 6 | Bq | 4 |
| 6 | Bm | 9 |
| 12 | Cz | 6 |

这里对原表做了两个修改：第一个修改将所有 y 的值以“A”开头的、x 值小于 2 的行中的 z 乘以 1.2；第二个修改将 x 的值分情况乘以不同的倍数：当 Y 以“A”开头时，乘以 2，当 Y 以“B”开头时，乘以 3，其他情况乘以 4。case 语句的用处很广，一些巧妙应用可以简化代码，十分方便，应多用活用。

【例 3-101】 删除数据。设程序名为 SASTJFX3_101.sas。

经常存在一些过时的数据，需要从表中删除，这时可以用 delete 语句实现。

```
data u3;
input x y $ z;
cards;
1 Ad 2
1 Ae 1
2 Afg 3
2 Bq 4
;
proc sql;
delete
from u3
where Y = 'Ad';
title "Deleted Data";
select *
from u3;
quit;
```

运行结果：

| Deleted Data | | |
|--------------|-----|---|
| x | y | z |
| 1 | Ae | 1 |
| 2 | Afg | 3 |
| 2 | Bq | 4 |

14. 数据列操作

对于数据列进行操作，需要使用 alter 语句。使用 alter 语句，可实现对表中列的添加、修改和删除操作。

【例 3-102】 添加列。设程序名为 SASTJFX3_102.sas。

```
data a1;
input x y $ z;
cards;
1 Ad 2
1 Ae 1
2 Afg 3
2 Bq 4
;
proc sql;
/* 添加一个数字列 */
alter table a1
add n num label = 'Number' format = 6.2;
/* 添加一个字符串列 */
alter table a1
add s char(15) label = 'String' format = $15.;
```

```
/* 为添加的列赋值* /
update a1 set n=x+4;
update a1 set s='String:'||y;
/* 输出结果* /
title "Altered Data";
select *
from a1;
quit;
```

运行结果:

Altered Data

| x | y | z | Number | String |
|---|-----|---|--------|------------|
| 1 | Ad | 2 | 5.00 | String:Ad |
| 1 | Ae | 1 | 5.00 | String:Ae |
| 2 | Afg | 3 | 6.00 | String:Afg |
| 2 | Bq | 4 | 6.00 | String:Bq |

添加列时，必须指明要添加列的名称是数字还是字符串，其他选项为可选。在添加字符串列时，最好指明字符串长度，如例子中的“char(15)”是声明添加一个长度为 15 个字符的字符串。

在 update 语句中，使用到“String:1||y”，意思是用字符串"String:"和变量 y 的值拼接成新的字符串。

【例 3-103】 修改列。设程序名为 SASTJFX3_103.sas。

使用 alter 下的 Modify 语句，只能改变列的长度、输入输出格式、标签。要改变列的名称，可以通过 DATA 步选项 RENAME = 来实现。

```
data a2;
input x y $ z;
cards;
1 Ad 2
1 Ae 1
2 Afg 3
2 Bq 4
;
proc sql;
/* 修改一个字符串列* /
alter table a2
    modify y char(15) label = 'StringY' format = $15.;
/* 输出结果* /
title "Altered Data";
select *
from a2;
quit;
```

运行结果:

Altered Data

| x | StringY | z |
|---|---------|---|
| 1 | Ad | 2 |
| 1 | Ae | 1 |
| 2 | Afg | 3 |
| 2 | Bq | 4 |

【例 3-104】 删除列。设程序名为 SASTJFX3_104. sas。

删除列使用 alter 下的 Drop 语句。

```
data a3;
  input x y $ z;
  cards;
1 Ad 2
1 Ae 1
2 Afg 3
2 Bq 4
;
proc sql;
  alter table a3
    drop z;
/* 输出结果* /
title "Altered Data";
select *
from a3;
quit;
```

运行结果:

| Altered Data | |
|--------------|-----|
| x | y |
| 1 | Ad |
| 1 | Ae |
| 2 | Afg |
| 2 | Bq |

3.4 SAS 数组介绍

3.4.1 概述

SAS 是一种面向数据集的编程语言,与其他编程语言不同,SAS 的一个变量并非只是一个值,而是代表一系列数据。对于一个变量的操作,等同于对一系列数据的操作。SAS 中的数组是针对数据集中变量进行的整合操作,对每一个数组元素的操作才是对数据的操作。

数组只能用在 SAS 数据步中。在处理 SAS 数据时,有时存在需要进行相同处理操作的一组或多组变量,利用数组可以简化操作。一般在 SAS 中,数组的作用如下。

- (1) 整理集合需要处理的变量。
- (2) 充当临时变量,作为计算中间量,不出现在数据集中。
- (3) 与 Retain 语句等效,可在数据步迭代读入数据的过程中始终保存数据。

3.4.2 Array 语法格式

语法格式:

ARRAY 数组名{下标} < \$ > < 长度 > < 数组元素 > < (元素初始值列表) >

选项说明:

| | | | |
|-----|-------------|---------|---------------|
| 数组名 | 规定数组名字 | 长度 | 规定字符数组每个元素的长度 |
| 下标 | 规定下标范围或元素个数 | 数组元素 | 给出组成数组的元素 |
| \$ | 声明数组为字符数组 | 元素初始值列表 | 规定数组中相应元素得初始值 |

注：语法中 < > 为非语法内容，表示该语句部分可省略，未加 < > 符号的语句为必写内容。

3.4.3 数组 Array 定义

要定义一个数组，首先必须给出确定的数组名，这里的数组名命名规则与 SAS 一般变量命名规则相同，同时要注意数组名不能与 SAS 中存在的函数重名。SAS 为了防止此类情况发生时出现的混乱，强制规定在后面出现该数组名时，统一作为数组处理，而非函数。要判断是否发生了重名，可以看执行程序的日志文档是否有提示。

数组与 SAS 函数重名：

```
data b;
  array abs (2) (1,2);
  put abs (2);
run;
```

执行程序，在日志中会出现以下提示。

NOTE: 数组 abs 与 SAS 提供的或用户定义的函数同名。该名称之后的括号按数组引用而非函数引用处理。

SAS 数组按变量类型可分为两类：一类是数值型，另一类是字符型。SAS 中不存在混合型数组，即在一个数组中，既包含数值变量又包含字符变量。

在定义数组时，数组的下标能告诉 SAS 程序，用户希望申请的数组维数大小信息，这在一般情况下是不可以省略的。下标不正确，经常导致所写程序出现问题，需要格外注意。数组下标可以通过以下几种方式实现定义。

| 下标 | 举例 |
|-------------------------|--------------------------------|
| 直接规定数组维数 | Array x {5,3} score1 - score15 |
| 规定数组下标范围 | Array x {2:4,3:6} |
| { * } ——根据后面的数组元素自动判断维数 | Array x { * } la lb lc; |
| 混合方式 | Array x {2:4,5} |
| | 这是一个 3 * 5 维数组 |

- 注：1. * 不可以作为混合方式使用。
2. 用范围规定维数时，前下标不大于后小标，如 x {4:3} 是不允许的，x {4:4}，x {4:5} 都是允许的。
3. 标注下标的符号可以是 {}, (), [], 三者均可。
4. 二维数组中，SAS 以行优先，按顺序将罗列的数组元素整合在数组中；高维数组中，同理，SAS 优先低位(定义中最后面一位)变化优先存储。例如定义“array x {5,3,2} s1-30”，先按顺序存储 x {1,1,1}，x {1,1,2} 后，开始存储 x {1,2,1}。

SAS 规定了 4 个特殊的变量名，可作为数组元素名称，实现特殊功能。

| 数组元素 | 说明 |
|-------------|----------------------------------|
| _NUMERIC_ | 将数据集中所有的数值变量作为数组元素 |
| _CHARACTER_ | 将数据集中所有的字符变量作为数组元素 |
| _ALL_ | 将数据集中所有的变量作为数组元素，需确保类型相同，否则会出错 |
| _TEMPORARY_ | 不能与下标为 { * } 共用。声明临时变量，不会出现在数据集中 |

1. 定义数值型数组和字符型数组

(1) 数值型数组

定义格式: ARRAY 数组名 {下标} <数组元素> <(元素初始值列表)>

例如:

```
array testa(0:2) RA RB RC (4,5,6);
array testb(3) English Chinese History;
array testc(*) English Chinese History;
array testd(3);
array teste(3) _numeric_;
```

说明: 数组 testa 规定了下标从 0 开始到 2, 并分别赋初值, 这样在 SAS 中使用从 0 维下标开始的数组, 如果用于大规模计算, 可提高程序运行效率。

数组 testb 和 testa 功能一样, 声明了 1 * 3 维数组, 由于没有赋初值, 在 SAS 数据集中都是缺失值。

testb 和 testd 的区别是, 数组中变量对应数据集中的变量名不同, testb 对应数据集中的 English、Chinese、History 三个变量, 如果数据集中已有此变量, 则直接引用; 如果没有则会生成新的以给定名称命名的变量。testd 对应数据集中的 testd1、testd2 和 testd3 三个变量, 编号由系统生成。

teste 将数据集中所有的数值变量集合在一起组成数组。

(2) 字符型数组

定义格式: ARRAY 数组名 {下标} \$ <长度> <数组元素> <(元素初始值列表)>

例如:

```
array testa(0:2) $ 10 RA RB RC ('a','b','c');
array testb(3) RA RB RE;
array testc(*) _character_ RF;
```

说明: 上面两个例子出现在同一个程序中, testb 依赖于 testa 的两个变量。在声明一个字符型数组时, "\$" 符号一般是必要的, 只有在特殊情况下才能省略掉, 否则会被 SAS 认为是数值型数组。在 testb 中, RA 和 RB 是数据集中已存在的两个字符型变量, 因此, SAS 判定 testb 为字符型数组, RE 虽然是新声明的变量, 但由于 SAS 数组不存在混合数值和字符的形式, 因此 RE 被 SAS 自动规定为字符型变量。testc 将数据集中存在的字符型变量 RA、RB、RC、RD 和 RE 集合在一起组成数组, 还声明了一个字符型变量 RF, 成为数组的最后一个元素。

例如, 变量长度设定:

```
data a;
input x1 $4. x2 $3.;
array t(10) $ 12 x1 -x10;
run;
```

例中, 定义了字符数组 t, 其中前两个变量在声明数组前已经定义了长度, 在数组中会保持其长度 4 和 3, 其余变量 x3 - x10 长度为数组规定值 12。

2. 特殊数组——隐含下标数组

SAS 中存在一种可以不用标明下标的数组, 与前面提到的数组不同的是, 它的下标是缺省

的，而数组元素不可缺，因此称为隐含下标数组。前面的数组称为显式下标数组。

类似于一般数组，语法格式如下：

```
ARRAY 数组名 <{下标}> <$> <长度> 数组元素 <(元素初始值列表)>
```

例如：array testa \$ RA RB RC('a','b','c');

说明：testa 的下标默认从 1 开始，在引用此类数组时，有一些方法是其独有的，请参看 3.4.5 节有关部分。

3. 临时数组

临时数组声明需要在数组元素中声明“_temporary_”，其他方面同一般数组没什么差别，需要注意的是不能用在下标维{*}和隐含下标数组中。

这里单独将临时数组拿出来，是由于临时数组特殊在它声明的变量可以不出现在数据集中。由于 SAS 赋予它相当于 RETAIN 语句的功能，声明一个临时数组，可保证在 DATA 步中一直保留，在任意时刻都可以引用，直到程序指定改变其值时，才会更新变量。因此，如果希望在程序中声明一些计算过程中的临时变量，而又不希望其出现在数据集中，可使用临时数组。

注意，在下面的声明方式中，_temporary_会出现在数据集中，不能实现上面所说的各项特点。

【例 3-105】 临时数组声明无效。设程序名为 SASTJFX3_105.sas。

```
data a;
input w h n $;
array t(2) w _temporary_;
cards;
61100 a
62120 b
63140 c
;
run;
```

执行程序后，数据集中共有 4 个变量，即 w、h、n 和名为 _temporary_ 的变量。

3.4.4 数组 Array 初始化

数组初始化，在定义数组及其元素的同时，给各个元素赋值，使其不再为缺省状态。初始化赋值根据数组类型不同，可以是数值，也可以是字符。这些初值必须在声明时用括号()引起来。

赋值时，可以直接将每个元素的初值罗列出来。

例如：

```
ARRAY x{10} x1-x10(1 2 3 4 5 6 7 8 9 10);
```

也可等价简写为 ARRAY x{10} x1-x10(1-10);

如果有连续几个变量相同，则可采用“n*(初值列表)”的形式。

例如：

```
ARRAY x{10} x1-x10(2*(1 2 3 4 5));
```

等价于 ARRAY x{10} x1-x10(1 2 3 4 5 1 2 3 4 5);

以下几种形式的结果一样，读者可以体会其中用法的灵活性：

```
ARRAY x{10} x1-x10(10*5);
```



```

ARRAY x{10} x1-x10(5*(5 5));
ARRAY x{10} x1-x10(5 5 3*(5 5) 5 5);
ARRAY x{10} x1-x10(2*(5 5) 5 5 2*(5 5));
ARRAY x{10} x1-x10(2*(5 2*(5 5)));

```

3.4.5 数组引用

对于数组的一般引用，在 SAS 中采用数组名 + 下标的方式。需要注意的是，在对数组元素进行引用时，如果定义的数组元素是数据集中的变量，则所有对数组变量的更改操作都等同于对数据集变量直接操作。

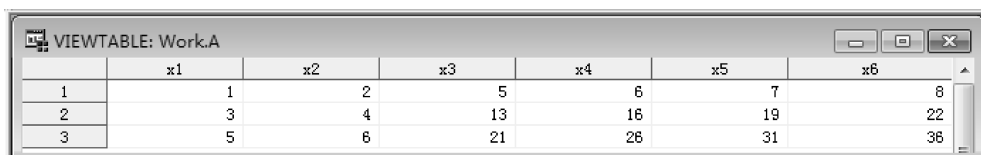
【例 3-106】 数组引用实例。设程序名为 SASTJFX3_106.sas。

```

data a(drop=i);
input x1 x2;
array t(6) x1-x6;
do i=3 to 6;
t(i)=x1*i+x2;
end;
cards;
1 2
3 4
5 6
;
run;

```

运行结果如图 3-7 所示。



| | x1 | x2 | x3 | x4 | x5 | x6 |
|---|----|----|----|----|----|----|
| 1 | 1 | 2 | 5 | 6 | 7 | 8 |
| 2 | 3 | 4 | 13 | 16 | 19 | 22 |
| 3 | 5 | 6 | 21 | 26 | 31 | 36 |

图 3-7 产生 6 个变量，各有 3 个观测

对于隐含下标数组的引用，除了一般方法外，还有两种独有的方法。

1. `_i_`代表下标

在隐含下标数组中，并不是说数组没有下标，默认为 `_i_`，这个也是可以更改的。看下面的例子：

【例 3-107】 设程序名为 SASTJFX3_107.sas。

```

data test;
input s1-s5;
array s s1-s5;
_i_=3;
s=s+2;
Do _i_=1 to 5;
s=s*20;
end;
cards;

```

```

1 1 1 1 1
1 1 1 1 1
1 1 1 1 1
;
run;

```

运行后，数据集变化如图 3-8 所示。

| | s1 | s2 | s3 | s4 | s5 |
|---|----|----|----|----|----|
| 1 | 20 | 20 | 60 | 20 | 20 |
| 2 | 20 | 20 | 60 | 20 | 20 |
| 3 | 20 | 20 | 60 | 20 | 20 |

图 3-8 TEST 数据集中包含 5 个变量，各有 3 个观测

默认下标可以改成任意变量名，还是前例中的程序，在声明隐含下标数组时，在下标列表中声明变量 num 之后可以用 num 代表其下标。两段程序运行结果完全相同。

```

data test;
input s1 - s5;
array s(num) s1 - s5;
num=3;
s=s+2;
do num=1 to 5;
s=s* 20;
end;
cards;
1 1 1 1 1
1 1 1 1 1
1 1 1 1 1
;
run;

```

2. Do Over 语句

该语句的作用是遍历隐含下标数组的所有元素，但仅限隐含下标数组可以使用，显式下标数组使用时会出错。

【例 3-108】 将数据中缺失值变为 0。设程序名为 SASTJFX3_108. sas。

```

data test;
input s1 - s5;
array s s1 - s5;
do over s;
if s=. then s=0;
end;
cards;
1 . 1 . 1
1 1 1 . 1

```

```
. 1 1 1 1
;
run;
```

3.4.6 有关数组的 SAS 函数

在引用使用数组时，尤其是多维数组在使用时经常容易弄错维数，或数组的上下限，而导致程序出错；有时，为了使程序更具有通用性，不希望在程序中具体写出某个数组维数，而希望程序能根据输入数组大小自行调整。例如在使用 do 循环时，对于一个具有 3 个元素的数组，循环表达式是“1 to 3”，而如果同样的程序，只是换了一下数组，元素个数也发生了变化，则需要手动修改循环表达式，此做法很烦琐，不利于程序的通用性。SAS 中的两个函数可以解决这一问题。

1. DIM 函数——求数组维数

语法：

```
DIM <n> (array-name)
DIM (array-name, bound-n)
```

选项说明：

| | |
|------------|-------------|
| n/bound-n | 指定多维数组第 n 维 |
| array-name | 数组名 |

【例 3-109】 使用 Dim 求多维数组维数。设程序名为 SASTJFX3_109.sas。

```
data test;
input s1 -s6;
array s(2,3) s1 -s6;
do i=1 to dim(s,1);
    do j=1 to dim2(s);
        if s(i,j) = . then s(i,j) = 0;
    end;
end;
cards;
1 . 1 . 1 1
1 1 1 . 1 1
. 1 1 1 1 1
;
run;
```

说明：例中，dim(s,1) 求数组第一维的维数，dim2(s) 求数组第二维的维数，分别采用两种不同的语法格式，功能上是等价的。

2. LBound 和 HBound——求数组上下界

SAS 中允许数组下标不从 1 开始，这虽然方便了用户使用，但也容易导致在使用中出错。当不确定数组上下界的情况下，可以借助 LBound 和 HBound 两个函数来解决这个问题。

语法如下：

求下界：

```
LBOUND <n> (array-name)
LBOUND (array-name, bound-n)
```

求上界：

```
HBOUND <n> (array-name)
HBOUND (array-name, bound-n)
```

语法结构同 DIM 类似，在此不再赘述。

更改例 3-109，使其更通用化

```
data test;
input s1 - s6;
array s(3:4,4:6) s1 - s6;
do i = lbound(s,1) to hbound(s,1);
    do j = lbound2(s) to hbound2(s);
        if s(i,j) = . then s(i,j) = 0;
    end;
end;
cards;
1 . 1 . 1 1
1 1 1 . 1 1
. 1 1 1 1 1
;
run;
```

说明：

例子中，双重循环也分别采用两种语法格式。这样，虽然数组在定义时，下标不是从 1 开始，但程序具有很强的鲁棒性，适用于各种类型二维数组。

3.5 SAS/IML 介绍

3.5.1 概述

在动态和交互式环境下，SAS/IML (Interactive Matrix Language) 是一种既强大又灵活的过程语言。用户可以通过 SAS/IML 模块立即获得程序运行结果，也可以将语句存储起来便于后续使用。SAS/IML 最基本的数据对象是矩阵。

SAS/IML 提供了许多内置模块来方便用户完成各种各样复杂的任务，如矩阵转置或者特征根求解等。在数据管理上，用户可以直接创建、读取以及更新数据集而避免使用数据步。同样，用户也可以根据需求编写自己的函数。

SAS/IML 软件具有以下特点。

- SAS/IML 软件是一种编程语言：SAS/IML 软件丰富的算术和字符表达式以及功能广泛的内置子程序可以使程序更简洁、更高效。SAS/IML 软件具有完备的控制语句，如 DO/END、START/FINISH、循环 DO、IF-THEN/ELSE、GOTO、LINK、PAUSE 和 STOP 语句等，为用户提供了程序执行控制及其模块化所需的全部命令。
- SAS/IML 软件是对矩阵进行操作：大多数编程语言的处理对象都是单一的数据元素，而 SAS/IML 软件是利用矩阵进行各式各样的计算。

- SAS/IML 软件具有丰富的函数和 CALL 子程序。
- SAS/IML 软件可以将运算符应用到这个矩阵：用户计算矩阵 A 与 B 对应元素的和可以直接使用表达式“A + B”来实现。
- SAS/IML 软件是交互的：用户可以选择立即执行程序或者将程序封装以后再执行。如果用户执行一个命令，将会立即看到结果。用户可以暂停一个模块的执行，然后键入需要补充的语句再继续运行。
- SAS/IML 软件是动态的：用户不需要声明一个矩阵的尺寸以及为其分配内存空间，SAS/IML 软件将自动完成这个工作。用户可以随时改变矩阵大小和类型，可以随时重置选项和更换模型，也可以同时打开多个文件或访问多个库。
- SAS/IML 软件可以处理数据：用户可以将数据集中全部观测或者有条件地选择部分观测存储到矩阵(也可以是多个向量)中，用户既可以创建新的数据集，也可以对已存在的数据集进行编辑或添加观测。
- SAS/IML 软件可以作图：SAS/IML 软件为用户提供了很多作图指令，便于用户通过图形找到数据之间的关系。

3.5.2 由矩阵标识创建矩阵

1. 矩阵的定义

SAS/IML 软件的基本运算对象是矩阵。这里矩阵的定义与矩阵代数中定义相同，是一个两维的(行 \times 列)数组，其中 $1 \times N$ 矩阵称为行向量， $M \times 1$ 矩阵称为列向量， 1×1 矩阵称为标量。SAS/IML 软件可以定义数值矩阵和字符矩阵。

矩阵的命名要考虑 SAS 编译器下的有效性，一般长度不能超过 8B，开头可以是字母或者下画线，后面可以用字母、数字以及下画线。矩阵的名字与它对应的值互相独立，也就是说用户随时都可以改变一个矩阵的尺寸及其内部的元素值，甚至对矩阵的类型(数字或字符)也可以随时调整。

矩阵可以用它的值标识出来。矩阵标识可以是单一的元素，也可以是行乘列的矩形形式，元素可以全是数值，也可以全是字符，但不能混合出现。

矩阵标识有多个元素时，需要用大括号将所有值括起来，并用逗号将行与行之间隔开，且每个行要有相同的元素数量。

- 数字可以用科学计数法的形式给出。
- 当没有大括号，或者要保持字符串的大小写，或者字符串内部有特殊字符时，用户必须用引号将字符串括起来。如果字符串内部自带引号，需要将其重复书写一次。特殊字符包括“?”、“=”、“*”、“:”、“()”。
- 符号“.”表示缺失的数值型元素。
- 中括号内的数值表示元素的重复因子。

2. 矩阵的创建

利用矩阵标识创建矩阵很简单，只需要将待输入的值用大括号括起来即可。当然，矩阵的创建方式有很多种，可以利用函数创建矩阵，也可利用 call 语句创建矩阵，还可利用赋值语句创建矩阵。下面介绍几种常见的矩阵创建方式。

【例 3-110】 标量的创建。

```
x=123;
x="saber";
x='chicken run';
```

例中定义了几种常见标量，写法是矩阵名字在左侧，元素值在右侧，中间用等号相连。当元素数量只有一个时，不必使用大括号。

【例 3-111】 数值矩阵的创建。

```
x={1 2 3 4 5 6};
y={1,2,3,4,5,6};
z={1 2,3 4,5 6};
```

矩阵元素有多个时需要使用大括号，行与行之间用逗号隔开，例中创建了三个数值矩阵，分别是行向量 x 、列向量 y 、 3×2 矩阵 z 。

【例 3-112】 字符矩阵的创建。

```
x={ab cDe,Fg hI};
y={"ab" 'cde','fg' "hi"};
```

输出结果：

| x | | y | |
|----|-----|----|-----|
| AB | CDE | ab | cde |
| FG | HI | fg | hi |

例中创建了两个 2×2 的矩阵。可以看到，如果字符串没有用引号括起来，那么 SAS 自动将其转化为大写。也就是说，如果要保持字符串的大小写，或者字符串中出现了特殊字符，那么引号是必不可少的。对于字符矩阵，元素的长度取决于所有元素中字符串长度最长的那个，而其他短的字符串将在后面自动添加空格。

【例 3-113】 含有重复元素的矩阵的创建。

```
x={{2} 1,[2] 2};
y={1 1,2 2};
```

例中两个语句创建矩阵的结果相同。对于含有位置相邻且元素值相同的矩阵，可以利用在中括号内部给出其重复次数，并在后面给出相应元素值来输入矩阵标识，从而达到简化矩阵创建的目的。

【例 3-114】 利用赋值语句创建矩阵。

```
x=2#y;
x=sqrt(y);
x=inv(y);
```

例中利用赋值语句完成矩阵的创建，等号右边分别用了计算表达式和函数表达式，对 IML 中的函数将在后面做详细的介绍。在表达式中使用运算符有常见的三种形式：一是前缀运算符 ($-A$)，二是中缀运算符 ($A * B$)，三是后缀运算符 (A')。在使用时，可以根据需要灵活变换，但要注意必须符合矩阵的运算法则。

上面已经强调过，如果矩阵元素有多个，大括号是必不可少的，同一矩阵内部的元素值类型必须相同(全是数值型或者全是字符型)，每一行必须有相同的元素个数。

下面给出一个较为完整的例子。

【例 3-115】 矩阵的创建和输出。设程序名为 SASTJFX3_115.sas。

```
proc IML;
reset print;
worktime = {6 7 8 8 8,
             7 7 7 8 7,
             8 8 8 8 8,
             7 8 6 8 8};
name = {'Archer', 'Lancer', 'Saber', 'Rider'};
print worktime [rowname = name];
run;
quit;
```

输出结果:

| worktime | 4 rows | | 5 cols | (numeric) | |
|----------|--------|--------|------------------------|-------------|---|
| 6 | 7 | 8 | 8 | 8 | |
| 7 | 7 | 7 | 8 | 7 | |
| 8 | 8 | 8 | 8 | 8 | |
| 7 | 8 | 6 | 8 | 8 | |
| name | 4 rows | 1 col | (character , size 6) | | |
| | | Archer | | | |
| | | Lancer | | | |
| | | Saber | | | |
| | | Rider | | | |
| worktime | | | | | |
| Archer | 6 | 7 | 8 | 8 | 8 |
| Lancer | 7 | 7 | 7 | 8 | 7 |
| Saber | 8 | 8 | 8 | 8 | 8 |
| Rider | 7 | 8 | 6 | 8 | 8 |

Reset print 语句使得程序一经提交，结果就被自动输出。可以看到前两部分输出结果包含矩阵的尺寸、类型以及元素值长度(当矩阵是字符型)。第三部分结果将两个矩阵结合起来，通过“rowname =”选项使得矩阵 worktime 的行标签被显示为矩阵 name。

【例 3-116】 利用冒号创建向量。设程序名为 SASTJFX3_116.sas。

```
x = 1:5;
y = 5:1;
z = 'x1 ':'x5 ';
xx = do(-1,1,0.5);
```

利用冒号可以很快捷地创建元素之间是递增或者递减关系的向量。如果要设定步长，可以使用 do 函数，例中第四行语句利用这个方法手动设定步长为 0.5。

3.5.3 矩阵操作

前一节介绍了矩阵创建的几种办法，有了矩阵之后再使用运算符就可以完成各种各样的矩阵运算操作。

1. 矩阵运算符的分类

前面介绍过，矩阵运算符可以大致分为以下三类。

(1) 前缀运算符。

(2)中缀运算符。

(3)后缀运算符。

首先要明确各个运算符的优先级，只有搞清楚了各个运算符的计算优先顺序，才能正确使用并实现矩阵的各种混合运算，参见表 3-25。

2. 矩阵运算符的应用

加号、减号、除号，元素相乘：+ - / #

加数和被加数可以有多种选择，如矩阵加矩阵、矩阵加标量、矩阵加向量。矩阵加矩阵就是简单的矩阵对应元素相加，即第一个矩阵 i 行 j 列的元素加上第二个矩阵 i 行 j 列的元素得到结果矩阵 i 行 j 列的元素，两个相加矩阵的尺寸要相同。

矩阵加标量的结果是由矩阵的每个元素均与这个标量相加所得到的。矩阵也可以与行向量或者列向量相加，结果矩阵是由向量与矩阵的每一行或者每一列分别相加所产生的。

【例 3-117】 “+”应用举例。设程序名为 SASTJFX3_117. sas。

```
proc IML;
reset print;
x={1 2 2,3 4 4};
y={2 5 8,4 9 1};
p=1:3;
q={1,2};
a=x+y;
b=x+p;
c=x+q;
run;
quit;
```

运行结果：

| | | | | | | | |
|---|--------|--------|------------|---|--------|--------|------------|
| a | 2 rows | 3 cols | (numeric) | c | 2 rows | 3 cols | (numeric) |
| | 3 | 7 | 10 | | 2 | 3 | 3 |
| | 7 | 13 | 5 | | 5 | 6 | 6 |
| b | 2 rows | 3 cols | (numeric) | | | | |
| | 2 | 4 | 5 | | | | |
| | 4 | 6 | 7 | | | | |

矩阵与行向量或者列向量求和时，该向量的长度一定要与矩阵的行或列相同。“-”作为减号时与加号用法基本相同。“/”的除数和被除数也可以有上面的三种情况，每种情况的计算方法与加号相同，但要注意作为除数的元素值不能为 0。元素相乘“#”计算的是矩阵元素之间的乘法，即两矩阵对应位置的元素相乘，要同矩阵与矩阵之间的乘法区分开。

比较运算符：< <= > >= = ^=

比较运算符可以用在矩阵与矩阵之间、矩阵与标量之间、矩阵与向量之间。结果矩阵的取值只有 0、1 之分，对应元素比较结果为真则取值为 1，对应元素比较结果为假则取值为 0。

如果矩阵元素为字符串时，系统会自动在短字符串的右侧填补空格，再进行比较；当用在条件语句时，结果矩阵必须所有元素均为 1，该条件才为真。

表 3-25 运算符优先级

| 优 先 级 | 运 算 符 |
|---------|------------------|
| I(最高) | ^ -(前缀) ## ** |
| II | * # < > > < / @ |
| III | + - |
| IV | // : |
| V | < < = > > = = ^= |
| VI | & |
| VII(最低) | |

【例 3-118】 比较运算符的应用。设程序名为 SASTJFX3_118. sas。

```
proc IML;
reset print;
x={1 2 2,3 4 4};
y={2 5 8,4 9 1};
p=1:3;
q={1,2};
a=(x>y);
b=(x<=p);
c=(x=q);
d=(x^=1);
run;
quit;
```

部分运行结果：

| | | | | | | | |
|---|--------|--------|-------------|---|--------|--------|-------------|
| a | 2 rows | 3 cols | (numeric) | b | 2 rows | 3 cols | (numeric) |
| | 0 | 0 | 0 | | 1 | 1 | 1 |
| | 0 | 0 | 1 | | 0 | 0 | 0 |

连接运算符：|| //

“||”可以将两个矩阵水平连接起来，两个矩阵必须有相同的行数，结果矩阵的行数即为原矩阵的行数，列数为两个矩阵的列数之和。

【例 3-119】 “||”的应用。设程序名为 SASTJFX3_119. sas。

```
proc IML;
reset print;
x={1 2 2,3 4 4};
y={2 5,4 9};
a=x||y;
run;
quit;
```

运行结果：

| | | | |
|---|--------|--------|-------------|
| a | 2 rows | 5 cols | (numeric) |
| 1 | 2 | 2 | 2 5 |
| 3 | 4 | 4 | 4 9 |

垂直连接运算符“//”与水平连接运算符用法类似，对矩阵要求有所不同。“//”要求两个矩阵列数相同，结果矩阵的列数即为原矩阵的列数，行数为两个矩阵的行数之和。

直乘运算符：@

结果矩阵是由两个矩阵直接相乘得到的(也叫克罗内克积)，新矩阵的行数为两个原矩阵的行数之积，列数为两个原矩阵的列数之积。

【例 3-120】 “@”的应用。设程序名为 SASTJFX3_120. sas。

```
proc IML;
reset print;
x={1 2 2,3 4 4};
y={2 5};
a=x@ y;
b=y@ x;
run;
quit;
```

运行结果：

| | a | 2 rows | 6 cols | (numeric) | |
|---|----|--------|--------|------------|----|
| 2 | 5 | 4 | 10 | 4 | 10 |
| 6 | 15 | 8 | 20 | 8 | 20 |
| | b | 2 rows | 6 cols | (numeric) | |
| 2 | 4 | 4 | 5 | 10 | 10 |
| 6 | 8 | 8 | 15 | 20 | 20 |

寻找元素最大值(最小值)：< > > <

参与比较的可以是两个矩阵，可以是矩阵与标量，也可以是矩阵与向量。两矩阵情形时，第一个矩阵的每个元素都会与第二个矩阵对应位置的元素进行比较，将最大值作为结果矩阵对应位置的元素；矩阵与标量情形时，矩阵的每个元素都会与该标量进行比较；矩阵与向量情形时，矩阵的每一行或者每一列都会与该向量进行比较。参与比较的元素也可以是字符串。

【例 3-121】 “< >”的应用。设程序名为 SASTJFX3_121. sas。

```
proc IML;
  reset print;
  x = {1 2 2,3 4 4};
  y = {2 5 8,4 9 1};
  p = 1:3;
  a = x < > y;
  b = x < > p;
  d = x < > 1;
  run;
  quit;
```

运行结果：

| a | 2 rows | 3 cols | (numeric) | d | 2 rows | 3 cols | (numeric) |
|---|--------|--------|------------|---|--------|--------|------------|
| | 2 | 5 | 8 | | 1 | 2 | 2 |
| | 4 | 9 | 4 | | 3 | 4 | 4 |
| b | 2 rows | 3 cols | (numeric) | | | | |
| | 1 | 2 | 3 | | | | |
| | 3 | 4 | 4 | | | | |

寻找最小值的运算符用法与上面介绍的基本相同，读者可以自行举例来加深对这个运算符的理解。

逻辑运算符：& | ^

参与运算的可以是矩阵与矩阵、矩阵与标量，也可以是矩阵与向量。

逻辑与“&”比较两矩阵每个对应位置的元素。当比较的两元素均非零时，得到的结果矩阵对应位置元素值为 1。

逻辑或“|”比较两矩阵每个对应位置的元素。当比较的两元素有一个非零时，得到的结果矩阵对应位置元素值为 1。当有一个运算因子为标量或向量时，其运算方法与前面介绍的加号、减号等情形相同。

逻辑非“^”对该矩阵每个元素进行逻辑运算。当元素值非零时，得到结果为 0；当元素值为零时，得到结果为 1。

【例 3-122】 逻辑运算符的应用。设程序名为 SASTJFX3_122. sas。

```
proc IML;
reset print;
x={1 2 2,0 4 4};
y={2 0 8,4 9 0};
p=0:2;
a=x&y;
b=x|p;
d=^x;
run;
quit;
```

运行结果：

| | | | | | | | |
|---|--------|--------|-------------|---|--------|--------|-------------|
| a | 2 rows | 3 cols | (numeric) | d | 2 rows | 3 cols | (numeric) |
| | 1 | 0 | 1 | | 0 | 0 | 0 |
| | 0 | 1 | 0 | | 1 | 0 | 0 |
| b | 2 rows | 3 cols | (numeric) | | | | |
| | 1 | 1 | 1 | | | | |
| | 0 | 1 | 1 | | | | |

矩阵相乘运算符：*

对参与运算的两矩阵尺寸要有严格的要求，即第一个矩阵的列数要与第二矩阵的行数相同，结果矩阵的行数与第一个矩阵的行数相同，结果矩阵的列数与第二个矩阵的列数相同。

【例 3-123】 “*”的应用。设程序名为 SASTJFX3_123. sas。

```
proc IML;
reset print;
x={1 2,0 4};
y={2,0};
a=x* y;
b=y'*x;
b=y'*x;
run;
quit;
```

运行结果：

| | | | | | | | |
|---|--------|-------|-------------|---|-------|--------|-------------|
| a | 2 rows | 1 col | (numeric) | b | 1 row | 2 cols | (numeric) |
| | | 2 | | | | 2 | 4 |
| | | 0 | | | | | |

元素乘方运算符：##

元素乘方对于矩阵与矩阵、矩阵与标量、矩阵与向量三种情形均适用。以矩阵与矩阵运算为例，第一个矩阵以第二个矩阵对应位置的元素为幂指数，计算得的结果作为结果矩阵对应位置的元素，如果第一个矩阵某一元素值为负，则与之相对应的幂指数须为整数。对于其中一个运算因子为标量或者向量时的处理方式与其他运算符相同，读者可参考前面的讲解。

【例 3-124】 “##”的应用。设程序名为 SASTJFX3_124. sas。

```
proc IML;
reset print;
x={1 2 ,3 8};
y={2 0.5};
```

```
a = x##y;
b = x##y`;
run;
quit;
```

运行结果：

| | | | | | | | |
|---|--------|-----------|-------------|---|-----------|-----------|-------------|
| a | 2 rows | 2 cols | (numeric) | b | 2 rows | 2 cols | (numeric) |
| | 1 | 1.4142136 | | | 1 | 4 | |
| | 9 | 2.8284271 | | | 1.7320508 | 2.8284271 | |

矩阵乘方运算符：**

矩阵乘方运算符只适用于矩阵与标量的情形。该矩阵以这个标量为幂指数运算得到结果，矩阵必须是方阵，标量是不小于 -1 的整数，数值太大则会导致计算精确度的问题。

【例 3-125】 “**”的应用。设程序名为 SASTJFX3_125. sas。

```
proc IML;
reset print;
x = {1 2 ,3 8};
a = x**2;
run;
quit;
```

运行结果：

| | | | |
|---|--------|--------|-------------|
| a | 2 rows | 2 cols | (numeric) |
| | 7 | 18 | |
| | 27 | 70 | |

当标量取值为 -1 时，等价于求该矩阵的逆矩阵。

转置运算符：`

矩阵转置在矩阵运算中十分常见，它是将原矩阵的行和列进行交换，原矩阵中第 i 行第 j 列的元素，在新矩阵中成为第 j 行第 i 列的元素，矩阵的尺寸也要进行相应的交换。

二目元素运算符参见表 3-26。

表 3-26 二目元素运算符

| 运算符 | 用途 | 运算符 | 用途 | 运算符 | 用途 | 运算符 | 用途 |
|-----|------|-----|------|-----|-------|----------|------|
| + | 加 | & | 逻辑与 | / | 除 | >= | 大于等于 |
| - | 减 | < | 小于 | < > | 元素最大值 | ^= | 不等于 |
| # | 元素相乘 | < = | 小于等于 | > < | 元素最小值 | = | 等于 |
| ## | 元素乘方 | > | 大于 | | 逻辑或 | MOD(m,n) | 余数 |

【例 3-126】 “`”的应用。设程序名为 SASTJFX3_126. sas。

```
proc IML;
reset print;
x = {1 2 3,3 8 5};
a = x`;
run;
quit;
```

运行结果：

| | | | |
|---|--------|--------|-----------|
| a | 3 rows | 2 cols | (numeric) |
| | 1 | 3 | |
| | 2 | 8 | |
| | 3 | 5 | |

3. 矩阵的下标

矩阵的下标([])是一种比较特殊的运算符，灵活地掌握矩阵下标的用法将会使 IML 编程代码更简单、高效。

一般形式为：operand[row ,colume]

- operand 通常是一个矩阵名，也可以是一个表达式。
- row 可以是矩阵的一行或几行，也可以是标量、向量或者表达式。
- colume 可以是矩阵的一列或几列，也可以是标量、向量或者表达式。

下标运算符可以完成以下工作。

- 引用矩阵的一个元素

有多种方式供引用矩阵的一个元素，用户可以使用行列下标直接给出元素的位置，或者只给出一个下标，系统会自动逐行进行搜索。

【例 3-127】 引用矩阵的一个元素。设程序名为 SASTJFX3_127. sas。

```
proc IML;
  reset print;
  a = {1 2 3, 3 8 5};
  x = {'x1', 'x2'};
  y = {'y1' 'y2' 'y3'};
  mattrib a rowname = x colname = y;
  a21_1 = a[2,1];
  a21_2 = a['x2', 'y1'];
  a21_3 = a[4];
  run;
  quit;
```

运行结果：

| | | | |
|-------|-------|-------|-----------|
| a21_1 | 1 row | 1 col | (numeric) |
| | | 3 | |
| a21_2 | 1 row | 1 col | (numeric) |
| | | 3 | |
| a21_3 | 1 row | 1 col | (numeric) |
| | | 3 | |

例中用了三种方式来引用矩阵的一个元素，第三种方式给出的是待引用元素逐行计数的标号。

- 引用矩阵的一行或一列

引用矩阵的一整行或列时，只需给出对应的行下标或者列下标，这里可以省略另一个下标，但是不能省略逗号。

【例 3-128】 引用矩阵的一行或一列。设程序名为 SASTJFX3_128. sas。

```

proc IML;
reset print;
a={1 2 3,3 8 5};
x={'x1','x2'};
y={'y1','y2','y3'};
mattrib a rowname=x colname=y;
ar2_1=a[2,];
ar2_2=a['x2',];
ac3_1=a[,3];
ac3_2=a[, 'y3'];
run;
quit;

```

运行结果:

| | | | | | |
|-------|-----------|--------|-------|-----------|-------|
| ar2_1 | 1 row | 3 cols | ac3_1 | 2 rows | 1 col |
| | (numeric) | | | (numeric) | |
| 3 | 8 | 5 | 3 | | |
| ar2_2 | 1 row | 3 cols | 5 | | |
| | (numeric) | | ac3_2 | 2 rows | 1 col |
| 3 | 8 | 5 | | (numeric) | |
| | | | 3 | | |
| | | | 5 | | |

例中用了两种方式引用矩阵的行或列：一种是直接给出行下标或者列下标，另一种是给出该行或列所对应的标签。

• 引用矩阵包含的子阵

我们可以直接给出指定的行和列，来引用矩阵的一个子阵，详见下面的例子。

【例 3-129】 引用矩阵包含的子阵。设程序名为 SASTJFX3_129.sas。

```

proc IML;
reset print;
a={1 2 3,3 8 5,4 7 0};
x={'x1','x2','x3'};
y={'y1','y2','y3'};
mattrib a rowname=x colname=y;
rows={1 3};cols={2 3};
b_1=a[{1 3},{2 3}];
b_2=a[rows,cols];
b_3=a[{'x1','x3'},{'y2','y3'}];
b_4=a[1:2,2:3];
run;
quit;

```

运行结果:

| | | | | | |
|-----|-----------|--------|-----|-----------|--------|
| b_1 | 2 rows | 2 cols | b_3 | 2 rows | 2 cols |
| | (numeric) | | | (numeric) | |
| 2 | 3 | | 2 | 3 | |
| 7 | 0 | | 7 | 0 | |
| b_2 | 2 rows | 2 cols | b_4 | 2 rows | 2 cols |

| (numeric) | | (numeric) | |
|-----------|---|-----------|---|
| 2 | 3 | 2 | 3 |
| 7 | 0 | 8 | 5 |

无论用怎样的方式来引用一个子阵，行列下标的顺序是不会改变的，先给出的必定是行下标，后给出的必定是列下标。

• 同时引用矩阵的多个元素

IML 中矩阵都是以行的方式存储的，所以理论上可以通过列出多个元素的位置来达到同时引用的目的；假设矩阵有 m 列，那么第一行第一个元素的位置是 1，第二行第一个元素的位置是 $m + 1$ ，第 n 行第一个元素的位置是 $(n - 1) * m + 1$ ，各行其他元素的位置都可以据此推算出来。

【例 3-130】 同时引用矩阵的多个元素。设程序名为 SASTJFX3_130.sas。

```
proc IML;
  reset print;
  a = {1 2 3, 3 8 5, 4 7 0};
  h = loc(a <= 4);
  b1 = a[h]';
  b2 = a[{1 2 3 4 7 9}]`;
run;
quit;
```

运行结果：

| | b1 | 1 row | 6 cols | (numeric) | |
|---|----|-------|--------|-----------|---|
| 1 | 2 | 3 | 3 | 4 | 0 |
| | b2 | 1 row | 6 cols | (numeric) | |
| 1 | 2 | 3 | 3 | 4 | 0 |

loc 函数返回的是满足条件的元素位置，有了这些位置，就可以顺利引用与之相对应的元素。

• 利用下标进行赋值

可以利用下标来引用矩阵的元素或子阵，再对其进行赋值，此时，下标要出现在等号的左侧。也可以用表达式给出下标，但表达式的结果必须为行向量或者列向量。

【例 3-131】 利用下标修改矩阵元素值。设程序名为 SASTJFX3_131.sas。

```
proc IML;
  reset print;
  a = {1 2 3, 3 8 5, 4 7 0};
  x = {'x1', 'x2', 'x3'};
  y = {'y1', 'y2', 'y3'};
  mattrib a rowname = x colname = y;
  a[2,1] = 0;
  a[3,] = {0 0 0};
  a[, 'y2'] = {0, 0, 0};
run;
quit;
```

运行结果：

| | | | | | | | |
|---|--------|--------|-----------|----|----|--------|--------|
| a | 3 rows | 3 cols | (numeric) | x2 | 0 | 8 | 5 |
| | y1 | y2 | y3 | x3 | 0 | 0 | 0 |
| | x1 | 1 | 2 | 3 | a | 3 rows | 3 cols |
| | x2 | 0 | 8 | 5 | | y1 | y2 |
| | x3 | 4 | 7 | 0 | | y3 | |
| a | 3 rows | 3 cols | (numeric) | x1 | 1 | 0 | 3 |
| | y1 | y2 | y3 | x2 | 0 | 0 | 5 |
| | x1 | 1 | 2 | 3 | x3 | 0 | 0 |

● 简化下标算符

利用简化下标算符可以得到一个降维的矩阵，既可降低行的维数，也可降低列的维数。

例如，想求一个矩阵各个列内的元素和(行的维数降为1)，用 `x[+,]` 语句：`+` 指定了求和并降维是发生在行，忽略了列下标使得列的维数保持不变，每一列的元素分别求和，得到结果是一个行向量。

可以使用运算符对行或列进行降维，也可以对二者同时操作，但是行的运算总是最先被执行的。

例如，表达式 `A[+,<>]` 返回的是所有列和中的最大值，而表达式 `A[,<>][+,]` 是对每行的最大值进行求和。简化下标算符见表 3-27。

已知 `A = {1 2 3, 3 8 5, 4 7 0}`，利用各种简化下标求得结果，见表 3-28。

表 3-27 简化下标算符

| 运算符 | 用途 | 运算符 | 用途 |
|-----|-----|------|-------|
| + | 加 | <: > | 最大值位置 |
| # | 元素乘 | >: < | 最小值位置 |
| < > | 求最大 | : | 求均值 |
| > < | 求最小 | ## | 求平方和 |

表 3-28 简化下标算符的应用举例

| 语 句 | 用 途 | 结 果 | 语 句 | 用 途 | 结 果 |
|--------------------------------|-------------|----------------------|---------------------------------|---------|----------------------|
| <code>A[{2 3},+]</code> | 矩阵第 2、3 行求和 | <code>{16,11}</code> | <code>A[,<: >]</code> | 行最大值位置 | <code>{3,2,2}</code> |
| <code>A[+,< >]</code> | 列和的最大值 | 17 | <code>A[>: <,]</code> | 列最小值位置 | <code>{1 1 3}</code> |
| <code>A[> <, +]</code> | 列最小值和 | 3 | <code>A[:,< >]</code> | 列平均的最大值 | 5.6667 |
| <code>A[,< >][+,]</code> | 行最大值和 | 18 | <code>A[,> <][: ,]</code> | 行最小值的平均 | 1.3333 |
| <code>A[,+][> <,]</code> | 行和的最小值 | 6 | <code>A[:]</code> | 所有元素均值 | 3.6667 |

4. 矩阵的混合表达式

使用 SAS/IML 时，用户可以使用含有多个矩阵运算符和运算对象的混合表达式，稍后将以几个表达式为例着重介绍混合表达式遵循的运算法则。

读者可以回顾表 9-1，从表中可以了解到，第一组的矩阵运算符有着最高的优先级，在混合表达式中将会被最优先计算，其他组的运算符将按优先级次序依次被计算；如果相邻的两个运算符有着相同的优先级，表达式将按从左到右的顺序计算，这一规律不适用于两运算符同时来自于第一组的情形；在括号中的表达式将会被优先计算。

● `a = x + y * z;`

由于乘号优先级高于加号，表达式优先计算 `y * z`，得出的结果再与矩阵 `x` 求和，最终结果赋值给矩阵 `a`。

● `a = x/y/z;`

上式将按照从左到右的运算顺序，先计算 `x/y`，得到的结果再与 `z` 比较，最终结果赋值给矩阵 `a`。

- $-x ** 2$;

上式将会先计算 x 的平方, 得到的结果再求其相反数。

- $A[i,j]$;

上式将会先对矩阵 A 求转置, 对结果矩阵寻找其第 i 行 j 列的元素。

- $a = x / (y / z)$;

上式先计算 y/z , 得到的结果作为除数与 x 进行除法运算, 最终结果赋值给 a 。

3.5.4 SAS/IML 编程语句

1. SAS/IML 基本编程语句

- IF-THEN/ELSE 条件语句

与 SAS 其他编程模块类似, 为了有条件地执行一项操作, 可以使用 IF-THEN/ELSE 语句, 供选择的操作分别出现在 THEN 子句和 ELSE 语句中。程序将先计算 IF 表达式, 如果结果是真, 则执行 THEN 后面的操作; 反之, 执行 ELSE 后面的操作, 没有 ELSE 的情况, 则直接执行下一条语句。

当条件表达式涉及矩阵时, 得到的结果一般是一个包含 0、1 及可能缺失值的矩阵; 如果该矩阵所有元素值均非零并且非缺失, 那么条件为真; 如果该矩阵任意一个元素为零, 那么条件为假。IF 语句可以嵌套使用, 理论上, 嵌套层数是没有限制的。

【例 3-132】 IF-THEN/ELSE 语句举例。设程序名为 SASTJFX3_132.sas。

```
proc IML;
reset print;
A={1 2 3,3 8 5,4 7 0};
B={0 1 2,2 7 4,3 6 0};
  if A[+, <>] >15 then flag1=1;
  if A<B then flag2=1;
else flag2=0;
else flag1=0;
run;
quit;
```

运行结果: $flag1 = 1$, $flag2 = 0$ 。

- DO 组语句

DO 组语句以 DO 语句作为开始, END 语句作为结束, 组内的一系列语句将被视为一个完整的单元来执行。针对一些程序上的要求, 有时必须使用 DO 组语句或者模块, 例如, LINK 和 GOTO 语句都必须出现在这样的单元中。

DO 组语句的嵌套层数也是没有限制的。将 DO 组语句结合条件语句使用, 在满足特定条件时, 可以一次性地执行多个操作。

【例 3-133】 DO 组语句举例。设程序名为 SASTJFX3_133.sas。

```
proc IML;
reset print;
A={1 2 3,3 8 5,4 7 0};
if A[+, <>] >15 then
  do;
    flag1=1;flag2=1;
  end;
```

```

else
  do;
    flag1=0;flag2=0;
  end;
run;
quit;

```

• 循环语句

利用 DO 语句可以完成循环操作，即重复执行一系列语句，直到满足特定条件结束循环为止。DO 语句的循环形式通常有下面四种形式。

(1) DO DATA 语句

一般形式是：DO DATA;

DATA 关键字声明了循环终止的条件是读到文件尾部，满足这个条件时，DO 循环立即终止，而其他形式的 DO 循环都是在循环单元的顶部或者尾部通过检验条件判断是否终止的。

(2) 循环 DO 语句

一般形式是：DO variable = start TO stop < BY increment >;

Variable 定义了循环变量，start 和 stop 分别定义了起始值和终止值，increment 指定了循环步长，默认是 1。Variable 从起始值开始循环，直到等于或超过了终止值时，循环终止。

(3) DO WHILE 语句

一般形式是：DO WHILE expression;

该语句每次循环开始时计算表达式的值。如果结果为零或缺失值，则终止循环。也就是说，如果第一次计算表达式的结果就为假，那么该循环单元将不被执行。

(4) DO UNTIL 语句

一般形式是：DO UNTIL expression;

该语句每次循环末尾计算表达式的值。如果结果为零或缺失值，则终止循环。也就是说，如果第一次计算表达式的结果就为假，那么该循环单元也将至少执行一次。

【例 3-134】 循环语句的应用举例。设程序名为 SASTJFX3_134. sas。

| | |
|---|---|
| <pre> proc IML; infile 'D:\my SAS files\test.txt'; do data; input xx; x = x//xx; end; sum1 = 0;sum2 = 0;sum3 = 0; do i = 1 to 9; sum1 = sum1 + x[i]; end; i = 1; do while (i < 10); </pre> | <pre> sum2 = sum2 + x[i]; i = i + 1; end; i = 1; do until (i > 9); sum3 = sum3 + x[i]; i = i + 1; end; print sum1 sum2 sum3; run; quit; </pre> |
|---|---|

运行结果: 39 39 39

例中利用 do data 语句读取文件中的数据，并垂直连接每一个数据得到一个列向量，后面三个 do 循环语句用了三种不同方式将这个列向量所有元素值进行求和，可以看到三个最终结果是相同的。

• 转移语句

在通常的编译过程，语句都是被逐条执行的，利用 GOTO 和 LINK 语句可以引导程序从一部分转移到另一部分。

LINK 语句的跳转是可以通过 RETURN 语句返回的，而 GOTO 语句的跳转是不可返回的。因此，LINK 语句提供了一种调用子程序的方式，子程序以 LABEL 作为开始，RETURN 作为结束。LINK 语句可以嵌套使用，理论嵌套层数没有限制。

需要注意的是：GOTO 语句和 LINK 语句在模块或者 DO 组语句中的使用是受限制的。转移语句指定跳转到的标记处也必须在当前的模块或者 DO 组语句中，尽管矩阵符号在模块之间是可以共享的，但语句标记不能共享，因此所有的 GOTO 语句或者 LINK 语句的标记也必须被一同放在 DO 语句或者模块中。当然，对于前述这种情形是有更好的替代方法的，应用之前讲到的循环语句是可以避免使用转移语句的。

【例 3-135】 转移语句的应用举例。设程序名为 SASTJFX3_135.sas。

```
proc IML;
x = 4;
do;
if x < 0 then goto negative;
y = sqrt(x);
print y;
stop;
negative:
print "Sorry, X is negative";
end;
run;
```

```
proc IML;
x = -4;
if x < 0 then print "Sorry, X is negative";
else
do;
y = sqrt(x);
print y;
end;
run;
quit;
```

上面两段程序实现了相同的功能，前一段程序用的是转移语句 GOTO，后一段程序用的是条件语句 IF-THEN/ELSE。可以看到，在使用转移语句时，GOTO 语句和标记 negative 必须出现在同一个 DO 组内部。

• 中断语句

常见的用来执行中断操作的语句有三种：PAUSE 语句、STOP 语句和 ABORT 语句。

这里以 PAUSE 语句为例简单介绍中断语句的用法，其一般形式为：

```
PAUSE <message> <* >;
```

PAUSE 语句可以实现以下几个功能：

- (1) 中断模块的执行。
- (2) 记忆中断运行的位置。
- (3) 输出指定的中断信息。
- (4) 进入模块环境的直接模式，该模式使用模块局部符号表，此时，用户可以输入更多的语句。

RESUME 语句可以使程序从最近的中断处继续执行。用户可以在 PAUSE 语句中指定输出的中断提示信息，在不指定的情况下，默认的提示信息为：

```
paused in module \ob x x x \obe
```

“x x x”表示的是包含中断的模块名称。如果不需要输出任何提示信息，语句形式可以是：

```
pause * ;
```

STOP 语句停止程序的执行并返回到直接模式，用户输入的新语句将被执行。如果中断是由 PAUSE 语句给出的，STOP 语句会清除所有的中断并返回到直接模式的执行。

在 ABORT 语句中断执行的同时直接退出 IML，功能类似于 QUIT 语句，但 ABORT 语句是可执行并且可编程的。在程序发生某种错误时，我们可以识别这个错误并利用 ABORT 语句直接退出 IML。只有在模块执行时，ABORT 语句才会执行，而 QUIT 语句无论何时都会立即执行并退出 IML。

2. 模块的定义和执行

IML 中应用模块有两个目的：

- (1) 创建子程序和函数，便于在程序任何地方随时对其进行调用。
- (2) 创建单独的符号表，可以定义属于模块的局部变量。

一个模块总是以 START 语句作为开始，以 FINISH 语句作为结束，既可以被视为子程序也可以被视为函数。能够返回一个参数的模块通常被称为函数，它与 IML 的内置函数没有本质差别，我们可以使用函数名来调用函数模块，而不是使用 CALL 或者 RUN 语句；非函数模块都可以被称为子程序，需要用 CALL 或者 RUN 语句来调用。

我们在模块外定义变量时，这个变量将被存储在全局符号表中。我们在模块外编程时，所用的变量都是来自于全局符号表。如果我们在 START 语句中定义了参数，那么系统将创建一个属于该模块的局部符号表，这个模块中使用的所有变量都将被存储在局部符号表中。局部符号表可以有多个，每个都对应一个含参数的模块，一个变量可以同时存在于一个全局符号表和多个局部符号表。局部表中的符号值都是临时的，也就是说，它们只存在于这个模块从执行开始到结束这个时间段，这个临时值是否会传递到对应的全局变量要取决于这个模块是怎样定义的。

• 模块的创建和执行

创建模块的一般形式为：

```
START <name> <( arguments )> <GLOBAL( arguments )>;
FINISH <name>;
```

默认模块名称为 MAIN，执行模块的一般形式有两种：

```
RUN name <( arguments )>;
CALL name <( arguments )>;
```

RUN 和 CALL 语句中的参数必须和待引用的模块参数相一致。当用户使用的模块名称与 IML 的内置子程序重名时，RUN 和 CALL 语句在解析名称上有顺序的差别，RUN 语句先解析为用户自定义模块，再解析为 IML 内置子程序或函数，CALL 语句先解析为 IML 内置子程序，再解析为用户自定义模块。因此，在面临这种情况时，我们一般使用 CALL 语句执行 IML 内置子程序，使用 RUN 语句执行用户自定义模块。

• 模块嵌套

模块之间可以互相嵌套。要保证嵌套的子模块必须完全包含在它所属的母模块中，每个模块是独立编译的。

【例 3-136】 模块的嵌套示例。设程序名为 SASTJFX3_136.sas。

```

start a;
reset print;
  start b;
  a = a + 1;
  finish b;
run b;
finish a;
run a;

```

例中，IML 先编译模块 A，编译过程中会遇到模块 B 的开始，系统会保存对模块 A 的编译状态，然后开始编译模块 B 直到遇到 FINISH 语句，模块 B 的编译结束后，再继续编译模块 A。

• 不含参数模块

当用户定义一个不含参数模块时，系统不会创建局部符号表，模块内定义的所有变量都是全局的，模块内外都可以引用这些变量，任何模块内的变量操作也都会影响到全局情况。

【例 3-137】 创建不含参数模块。设程序名为 SASTJFX3_137.sas。

```

proc IML;
x=1;y=5;
z="temp";
start test;
  a=x+y;y=y*2;
  z="IML";
finish;
run test;
print x y z a;
quit;

```

运行结果：

| x | y | z | a |
|---|----|-----|---|
| 1 | 25 | IML | 6 |

运行之后，x 是 1，y 是 25，z 是“IML”，三者都是全局变量，a 是模块内部创建的变量，对应值是 6，也是全局变量。

• 含参数模块

一般来说，含参数模块有以下几个特点：

- (1) 可以用变量名的方式指定参数。
- (2) 如果指定了多个参数，需要用逗号将其隔开。
- (3) 如果既有输入变量又有输出变量，最好把输出变量列在前面。
- (4) 当使用 RUN 或者 CALL 语句调用模块时，参数可以是名称或表达式，然而，在参数用于输出结果时，必须使用变量名，而不能使用表达式。

【例 3-138】 创建含参数模块。设程序名为 SASTJFX3_138.sas。

```

proc iml;
a=1; b=2;
c=3; d=9;
start test(x,y);
p=x+y; q=y-x;
y=10; c=100;
finish;
run test(a,b);

```

```
print a b c d;
quit;
```

运行结果:

| a | b | c | d |
|---|----|---|---|
| 1 | 10 | 3 | 9 |

a 值保持不变, b 值因为参数传递由 2 变成了 10, c 在模块内部被赋值 100, 但是这只改变了 c 在局部符号表中的取值, 并不会影响到它在全局符号表中的取值, d 依然取值为 9。

当使用 RUN 或 CALL 语句调用模块时, 每一个参数值都从全局符号表传递到局部符号表。例中, 首先在模块外定义四个变量(a、b、c、d), 显然, 它们的值将被存放在全局符号表中, 然后定义一个含有两个参数(X, Y)的模块 test, 这一操作使系统为模块 test 创建了一个局部符号表, 所有在模块内使用的变量都将被放在局部符号表中; 例子中的变量 a、b、d 只存在于全局符号表中, 变量 X、Y、P、Q 只存在于局部符号表中, 而变量 C 同时存在于局部和全局符号表中; 当使用 RUN 语句调用模块 test(a,b)时, 全局符号表中的 A 值传递到局部符号表中的 X, 类似地, 全局符号表中的 B 值传递到局部符号表中的 Y; 这里 c 不是参数, 因此, 全局符号表中的 c 值和局部符号表中的 C 值不具有一致性; 当模块结束执行时, 局部符号表中的 X、Y 值会传回到全局符号表中的 a、b。

● 创建函数模块

函数是可以返回一个值的特殊模块。一般, 对通过参数列表返回变量的数量也是没有限制的。对于一个函数模块来说, RETURN 语句是必需的。调用函数模块时可以用赋值语句, 函数模块在符号表上的特性与一般参数模块相同。

【例 3-139】 自定义函数模块。设程序名为 SASTJFX3_139.sas。

```
proc iml;
start maxx(x,y);
if x <= y then return(y);
else return(x);
finish;
a=4;b=8;
c=maxx(a,b);
d=maxx(maxx(a+c,b),maxx(c-a,c+a));
print c d;
quit;
```

运行结果:

| c | d |
|---|----|
| 8 | 12 |

例中揭示了函数之间是可以互相调用的, 而且调用时是可以传递表达式值的。函数模块遵从这样一个解析过程, 先解析成 IML 内置函数, 再解析成用户自定义函数, 最后解析成 SAS 数据步函数。

● 使用 GLOBAL 子句

对于函数参数模块, 模块内创建的局部变量和模块外创建的全局变量没有任何联系。然而, 我们是能够在模块中定义全局变量的, 使用 GLOBAL 子句可以使指定变量在局部和全局符号表中都能被引用。

【例 3-140】 global 子句使用实例。设程序名为 SASTJFX3_140.sas。

```
proc iml;
  a=1; b=2;
  c=3; d=9;
  start test(x,y) global(c);
  p=x+y; q=y-x;
  y=10; c=100;
  finish;
  run test(a,b);
  print a b c d;
quit;
```

运行结果:

| | | | |
|---|----|-----|---|
| a | b | c | d |
| 1 | 10 | 100 | 9 |

例中使用了 global 子句将变量 c 定义成全局性的,使得变量 c 的值在全局和局部符号表间被共享,因此例中在模块内对 c 值的修改对于模块外一样生效。

由于每个模块都有自己的局部符号表,因此一个全局符号表和多个局部符号表并存的情形是很常见的。一个变量可以独立地存在于这些表中,也可以在全局表和几个局部表间被共享(使用 global 子句)。

● 模块的储存

用户可以使用 STORE 和 LOAD 语句进行模块的存储和重载。

```
STORE MODULE = name;
LOAD MODULE = name;
```

可以用 show 语句来显示已经存储的模块名称。

```
Show storage;
```

3. IML 中命令语句

IML 中命令语句常用来完成一些指定的系统操作,例如存储及重载矩阵或模块,完成特殊的数据处理请求。表 3-29 列出了常用命令语句及其功能。

表 3-29 常用命令语句及其功能

| 命 令 | 形 式 | 功 能 |
|---------|---|-----------------|
| FREE | FREE matrices; | 释放矩阵占用内存,增加可用空间 |
| LOAD | LOAD < MODULE = (module-list) > < matrix-list >; | 从存储库中加载矩阵或者模块 |
| MATTRIB | MATTRIB name < ROWNAME = row-name > < COLNAME = col- umn-name > < LABEL = label > < FORMAT = format >; | 把矩阵和打印属性联系起来 |
| PRINT | PRINT < matrices > < (expression) > < "message" > < pointer-controls > < [options] >; | 打印矩阵或信息 |
| RESET | RESET < options >; | 同时设置多个系统选项 |
| REMOVE | REMOVE < MODULE = (module-list) < matrix-list >; | 从存储库中移除矩阵或者模块 |
| SHOW | SHOW operands; | 发出显示系统信息的请求 |
| STORE | STORE < MODULE = (module-list) > < matrix-list >; | 将矩阵或者模块存储到库中 |

这些命令在 IML 中有很重要的应用地位,通过这些语句可以控制输出显示与否(RESET PRINT、RESET NOPRINT 或者 MATTRIB),可以获得各种系统信息(SHOW SPACE、SHOW STORAGE 或者 SHOW ALL)。

如果遇到了可用空间短缺的情形，可以用命令将矩阵存储到库中，然后释放这些矩阵占用的内存，这样就可以增多现阶段的可用空间，而释放的矩阵可以在稍后必要时再进行重载。

FREE 语句释放了指定矩阵所占用的内存空间，该矩阵将失去赋值，行列数目均为 0。FREE 语句主要应用在数据量比较大而内存相对紧张的情形。如果想要释放大部分矩阵，而只保留少数矩阵，可以用 FREE/ <matrices> 语句，括号内填写需要保留的矩阵名。

LOAD 语句可以把当前库存中的模块或矩阵加载到当前的工作空间中，结合特殊指令 _ALL_ (module = _all_) 可以加载全部的矩阵(模块)。

MATTRIB 语句将打印属性和矩阵联系起来。每个矩阵的行或列标签都可以用矩阵来存储，必要时再将标签矩阵赋给行名或列名。

PRINT 语句可以打印矩阵或信息，可以临时地将打印信息和矩阵联系起来。如果没有足够的空间在同一页中输出所有矩阵，那么超出的矩阵将会被放到下一页中。对于列数超出一页显示容量的矩阵，超过的列数会被折叠起来，并用冒号“:”作为延长线来代替。

RESET 语句主要用来设置一些处理选项，具体有哪些选项及其代表的意义请查阅说明书。

REMOVE 语句可以从当前库中移除指定的矩阵或模块，或者将二者同时移除，结合 _ALL_ 指令可以移除全部的矩阵或模块。

SHOW 语句可以输出系统信息，例如变量名及其属性、激活的数据集、打开的文件、存储在库中的矩阵和模块等，更多内容请查阅说明书。

STORE 语句可以把当前工作空间中的模块或矩阵存储到当前库中，结合特殊指令 _ALL_ (module = _all_) 可以存储全部的矩阵(模块)。

3.5.5 IML 中常用函数

本章只介绍几个常见的 IML 函数，有些函数(如 MIN、MAX 等)虽然常见但其用法很简单，所以没有加以介绍，更多复杂函数或者有专业需求的函数请查阅帮助文档。

1. 矩阵生成函数

IML 中有很多内置的矩阵生成函数，常用的包括 BLOCK 函数、DESIGN 函数、I 函数、J 函数、SHAPE 函数。

● BLOCK 函数

BLOCK 函数引用的一般形式为：

```
BLOCK(matrix1, <matrix2, ..., matrix15>);
```

该函数利用给定的参数矩阵创建了一个块状对角矩阵，括号内最多可一次性选取 15 个参数矩阵。这些矩阵基于对角线连接起来得到了一个新矩阵。

例如，语句 block(A,B,C)的结果是

$$\begin{bmatrix} A & 0 & 0 \\ 0 & B & 0 \\ 0 & 0 & C \end{bmatrix}$$

【例 3-141】 创建块对角矩阵。设程序名为 SASTJFX3_141.sas。

```
proc IML;                                C=block(A,B);
reset print;                             run;
A={1 1};                                quit;
B={1 2,3 4};
```


运行结果：

| C | 3 rows | 4 cols | (numeric) |
|---|--------|--------|-------------|
| 1 | 1 | 0 | 0 |
| 0 | 0 | 1 | 2 |
| 0 | 0 | 3 | 4 |

● DESIGN 函数

DESIGN 函数引用的一般形式为：

```
DESIGN (column-vector)
```

括号内是一个数值型的列向量。

DESIGN 函数根据参数列向量创建一个用户自定义的 0、1 组合矩阵，参数中有几个不同的取值就会产生几列，每一列 1 的数量由列向量中对应该列的元素出现次数来决定，其余位置由 0 填充。

【例 3-142】 创建 0、1 组合矩阵。设程序名为 SASTJFX3_142. sas。

```
proc IML;                                C=design(A);
reset print;                             C=design(B);
A={2,3,5};                               run;
B={1,-1,3,-1};                           quit;
```

运行结果：

| C | 3 rows | 3 cols | (numeric) | C | 4 rows | 3 cols | (numeric) |
|---|--------|--------|-------------|---|--------|--------|-------------|
| | 1 | 0 | 0 | | 0 | 1 | 0 |
| | 0 | 1 | 0 | | 1 | 0 | 0 |
| | 0 | 0 | 1 | | 0 | 0 | 1 |
| | | | | | 1 | 0 | 0 |

可以看到列的数量取决于列元素取值的不同个数，而与大小没有任何关系，元素值具体对应某一列要根据元素实际取值的大小来排列，例中参数设为 B 向量时，参数值 -1 对应的第一列，1 对应的第二列，3 对应的第三列；结果矩阵行数由列向量个数决定的。

DESIGNF 函数与之功能类似，但是得到的结果矩阵会比 DESIGN 函数得到的结果矩阵少最后一列，然后用其他列减去去掉的最后一列对应位置元素作为最终结果矩阵。

【例 3-143】 利用 DESIGNF 创建矩阵。设程序名为 SASTJFX3_143. sas。

```
proc IML;                                C1=designf(A);
reset print;                             C1=designf(B);
A={2,3,5};                               run;
B={1,-1,3,-1};                           quit;
```

运行结果：

| C1 | 3 rows | 2 cols | (numeric) | C1 | 4 rows | 2 cols | (numeric) |
|----|--------|--------|-------------|----|--------|--------|-------------|
| | 1 | 0 | | | 0 | 1 | |
| | 0 | 1 | | | 1 | 0 | |
| | -1 | -1 | | | -1 | -1 | |
| | | | | | 1 | 0 | |

• I 函数

I 函数引用的一般形式为：

```
I (dimension)
```

括号内参数用来定义矩阵尺寸。

I 函数创建的是一个单位矩阵，即对角线上元素值全为 1，其他位置元素值全为 0，dimension 给出的是行和列的数目，取值必须是整数并且大于等于 1。

【例 3-144】 创建单位矩阵。设程序名为 SASTJFX3_144.sas。

```
proc IML;
reset print;
A = I(3);
run;
quit;
```

运行结果：

| A | 3 rows | 3 cols | (numeric) |
|---|--------|--------|-------------|
| | 1 | 0 | 0 |
| | 0 | 1 | 0 |
| | 0 | 0 | 1 |

• J 函数

J 函数引用的一般形式为：

```
J(nrow <, ncol <, value >)
```

J 函数创建了一个 NROW 行、NCOL 列，并且所有元素均等于 VALUE 的矩阵。如果省略列参数，则默认列参数等于行参数。如果省略参数值的定义，则默认值为 1。

【例 3-145】 利用 J 函数创建矩阵。设程序名为 SASTJFX3_145.sas。

```
proc IML;
reset print;
A = J(2);
B = J(2,3,'saber');
run;
quit;
```

运行结果：

| A | 2 rows | 2 cols | (numeric) | B | 2 rows | 3 cols | (character, size 5) |
|---|--------|--------|-------------|---|--------|--------|-----------------------|
| | 1 | 1 | | | saber | saber | saber |
| | 1 | 1 | | | saber | saber | saber |

• SHAPE 函数

SHAPE 函数引用的一般形式为：

```
SHAPE(matrix <, nrow <, ncol <, pad-value >)
```

SHAPE 函数根据给定的参数构造出指定尺寸的各种样式矩阵，这个函数的使用方法简单概括如下。

(1) 如果只给定行参数, 那么列参数将由参数矩阵的元素总数除以行参数得到, 元素个数必须能够被行参数整除, 否则将会报错。

(2) 如果同时给定参数矩阵、行参数和列参数, 但没有给出填充值, 那么系统将会逐行读取参数矩阵的元素值, 直到达成预期的元素数量。必要时, 系统会重复读取参数矩阵以达到参数要求。

(3) 如果给定了填充值, 系统会先读取参数矩阵中的元素值, 再将多余的位置写入填充值作为最终的结果矩阵。

【例 3-146】 利用 SHAPE 函数创建矩阵。设程序名为 SASTJFX3_146.sas。

```
proc IML;
  reset print;
  A = shape(66,3,2);
  B = shape({9 8 7,6 5 4},2,2);
  C = shape({9 8 7,6 5 4},3);
  D = shape({11,22},2,3,33);
run;
quit;
```

运行结果:

| | | | | | | | |
|---|--------|--------|-------------|---|--------|--------|-------------|
| A | 3 rows | 2 cols | (numeric) | C | 3 rows | 2 cols | (numeric) |
| | 66 | 66 | | | 9 | 8 | |
| | 66 | 66 | | | 7 | 6 | |
| | 66 | 66 | | | 5 | 4 | |
| B | 2 rows | 2 cols | (numeric) | D | 2 rows | 3 cols | (numeric) |
| | 9 | 8 | | | 11 | 22 | 33 |
| | 7 | 6 | | | 33 | 33 | 33 |

对于矩阵 A, 参数 66 可以看成是一个标量, 而不能看成填充值, 填充值是列在参数后面的; 矩阵 B 是 2×2 的方阵, 它是由参数矩阵逐行读取直到 4 个元素为止; 矩阵 C 只给定了行参数, 而参数矩阵元素个数为 6, 据此得到列参数为 2, 因此结果矩阵是 3×2 的; 矩阵 D 给定了列向量作为参数, 而结果矩阵要求是 2×3 的, 当不给定填充值时, 系统将会重复读取这个列向量直到元素个数达到要求, 当给定填充值时, 第一次读取完列向量之后余下的位置都写入填充值。

• REPEAT 函数

REPEAT 函数引用的一般形式为:

```
REPEAT(matrix, nrow, ncol)
```

REPEAT 函数将参数矩阵重复 $nrow \times ncol$ 次得到结果矩阵, 参数矩阵既可以是数值的, 也可以是字符的。

【例 3-147】 参数矩阵重复指定次数创建新矩阵。设程序名为 SASTJFX3_147.sas。

```
proc IML;
  reset print;
  A = shape(66,2,2,55);
  B = repeat(A,2,2);
run;
quit;
```

运行结果：

| B | 4 rows | 4 cols | (numeric) |
|----|--------|--------|-------------|
| 66 | 55 | 66 | 55 |
| 55 | 55 | 55 | 55 |
| 66 | 55 | 66 | 55 |
| 55 | 55 | 55 | 55 |

2. 矩阵查询函数

• LOC 函数

LOC 函数引用的一般形式为：

LOC (matrix)

LOC 函数创建了一个 $1 \times n$ 的行向量，这里的 n 是参数矩阵中非零元素的个数，缺失值默认视为 0，结果行向量中的值是参数矩阵中非零元素的位置(逐行查询)。LOC 函数可以应用在寻找满足某条件的子矩阵。

【例 3-148】 寻找满足指定条件的子矩阵。设程序名为 SASTJFX3_148.sas。

```
proc iml;
reset print;
A = {-2,1,0,3};
B = loc(A > 0);
C = A[B,];
run;quit;
```

运行结果：

| B | 1 row | 2 cols | (numeric) | C | 2 rows | 1 col | (numeric) |
|---|-------|--------|-------------|---|--------|-------|-------------|
| | 2 | 4 | | | | 1 | |
| | | | | | | 3 | |

• RANK 函数

RANK 函数引用的一般形式为：

RANK (matrix)

RANK 函数运行结果得到一个新矩阵，该矩阵包含了对应元素的排秩，对于相等的矩阵元素，它们的秩是几个元素对应的顺序秩(而非 RANKTIE 函数求平均值)。

【例 3-149】 对矩阵元素排序。设程序名为 SASTJFX3_149.sas。

```
proc iml;
reset print;
A = {-2 1 1, 0 3 0.5};
B = rank(A);
C = A;
A[B] = C;
run;quit;
```

运行结果：

| A | 2 rows | 3 cols | (numeric) | A | 2 rows | 3 cols | (numeric) |
|---|--------|--------|-------------|---|--------|--------|-------------|
| | -2 | 1 | 1 | | -2 | 0 | 0.5 |
| | 0 | 3 | 0.5 | | 1 | 1 | 3 |

例子首先利用 RANK 函数根据矩阵 A 中的元素大小给出对应的排秩，然后再将元素按对应的秩重新进行排布，即完成了矩阵元素排序。

- CHOOSE 函数

CHOOSE 函数引用的一般形式为：

```
CHOOSE(condition, result-for-true, result-for-false)
```

CHOOSE 函数检查第一个参数的每个元素是否满足指定的条件式，当条件式为真时，函数返回第二个参数对应位置元素；当条件式为假时，函数返回第三个参数对应位置元素。

【例 3-150】 利用 choose 函数实现 if-else 语句功能。设程序名为 SASTJFX3_150.sas。

```
proc iml;
reset print;
x = {5 7, 3 9};
y = {8 2, 4 6};
r = choose(x > y, x, y);
run;quit;
```

运行结果：

| r | 2 rows | 2 cols | (numeric) |
|---|--------|--------|-----------|
| | 8 | 7 | |
| | 4 | 9 | |

例中当 x 的元素大于 y 的元素时，则将 x 的元素赋给 r，反之，将 y 的元素赋给 r，实际上等同于用 if-else 语句逐个比较两矩阵元素，选择大者赋给结果矩阵。

- UNION 函数

UNION 函数引用的一般形式为：

```
UNION(matrix1 <, matrix2, ..., matrix15 >)
```

该函数返回结果是一个排序好的行向量，是由各参数矩阵所有不相同的元素组成的，参数矩阵最多不超过 15 个，但类型必须相同，可以统一是数值型的，也可统一是字符型的。如果是字符型的，元素长度由参数矩阵中最长的元素来决定，其他短的元素将在右侧用空格填补，UNIQUE 函数功能与之相同。

【例 3-151】 求出各矩阵全部不相同元素。设程序名为 SASTJFX3_151.sas。

```
proc iml;
reset print;
a = {1 1 1, 1 2 4, 1 3 9};
b = {1, 2, 2};
c = union(a, b);
run;quit;
```

运行结果：

| c | 1 row | 5 cols | (numeric) |
|---|-------|--------|-----------|
| 1 | 2 | 3 | 4 |

3. 数学函数

- APPLY 函数

APPLY 函数引用的一般形式为：

```
APPLY(modname, argument1 <, argument2, ..., argument15 >)
```

APPLY 函数将用户自定义的模块应用到参数矩阵的每一个元素上，并返回一个最终的结果矩阵。第一个参数是模块名，该模块必须在 APPLY 函数执行前定义好，而且必须是有返回值的函数模块。

随后的参数都是要传递给模块的，因此这些参数与函数模块参数要有相同的尺寸大小，参数个数不能超过 15 个。结果矩阵的每一个元素都是由参数矩阵对应位置元素经函数模块计算得到的，参数矩阵既可以是数值的，也可以是字符的。

【例 3-152】 利用 APPLY 函数实现矩阵元素相乘。设程序名为 SASTJFX3_152.sas。

```
proc IML;
  start AXB(a,b);
  c = a* b;
  return(c);
  finish;
  reset print;
  a = {2 3,8 5};
  b = {3 7,4 9};
  mod = {AXB};
  c = apply(mod,a,b);
  run;
  quit;
```

运行结果：

| | | | |
|---|--------|--------|-------------|
| c | 2 rows | 2 cols | (numeric) |
| | 6 | 21 | |
| | 32 | 45 | |

• DET 函数

DET 函数引用的一般形式为：

```
DET(square-matrix)
```

DET 函数可以计算一个方阵的行列式。行列式就是一个单一数值，可以由矩阵特征根乘积得到。原理是对方阵进行 LU 分解，再求得对角线元素乘积(具体可查阅相关书籍)。

【例 3-153】 利用 DET 函数求方阵行列式。设程序名为 SASTJFX3_153.sas。

```
proc iml;
  reset print;
  a = {1 1 1,1 2 4,1 3 9};
  b = det(a);
  run;
  quit;
```

运行结果：

| | | | |
|---|-------|-------|-------------|
| b | 1 row | 1 col | (numeric) |
| | | 2 | |

• DIAG 函数

DIAG 函数引用的一般形式为：

DIAG(argument)

如果参数是一个方阵(矩阵时必须是方阵),那么该函数返回一个只包含参数矩阵对角元素的新矩阵,所有非对角元素都将被置为0;如果参数是一个向量,该函数将以参数的每个值作为对角元素创建一个新矩阵,所有非对角元素都将被置为0。

【例 3-154】 求矩阵对角元素。设程序名为 SASTJFX3_154.sas。

```
proc iml;
reset print;
x = {5 7, 3 9};
r = diag(x);
run;quit;
```

运行结果:

| | | | |
|---|--------|--------|-----------|
| r | 2 rows | 2 cols | (numeric) |
| | 5 | 0 | |
| | 0 | 9 | |

- INV 函数

INV 函数引用的一般形式为:

INV(matrix)

INV 函数可以求得指定参数矩阵的逆矩阵,前提是这个参数矩阵必须是可逆的。一个矩阵的逆矩阵和它本身的乘积是一个单位阵,这一性质可以应用于线性方程组的求解。

【例 3-155】 求指定矩阵的逆矩阵。设程序名为 SASTJFX3_155.sas。

```
proc iml;
reset print;
A = {2 2, 2 3};
B = inv(A);
r = B* A;
run;quit;
```

运行结果:

| | | | | | | | |
|---|--------|--------|-----------|---|--------|--------|-----------|
| B | 2 rows | 2 cols | (numeric) | r | 2 rows | 2 cols | (numeric) |
| | 1.5 | -1 | | | 1 | 0 | |
| | -1 | 1 | | | 0 | 1 | |

- SOLVE 函数

SOLVE 函数引用的一般形式为:

SOLVE(A, B)

该函数解决了形如 $AX = B$ 的线性方程组求根问题,矩阵 A 必须是方阵并且非奇异。实际上,根据前面的矩阵求逆函数,对于上述方程组我们有 $X = \text{INV}(A) * B$,这与 SOLVE 函数达到的目的是相同的,但我们推荐使用 SOLVE 函数,因为后者计算更加精确高效。

【例 3-156】 线性方程组求根。设程序名为 SASTJFX3_156.sas。

```
proc iml;
reset print;
a = {1 1 1, 1 2 4, 1 3 9};
b = {1, 2, 2};
X = solve(a, b);
run; quit;
```

运行结果：

```

X          3 rows      1 col      (numeric)
          -1
           2.5
          -0.5
```

线性方程组求根在实际应用中十分常见，SOLVE 函数为我们解决这类问题提供了很多方便。

3.5.6 IML 中数据集的操作

SAS/IML 为我们提供了很多语句来实现数据集和矩阵之间的相互传递。可以通过多种方式将数据集中的观测写入到矩阵，既可以选择全部观测，也可以选择部分观测。可以将每一个观测对应矩阵中的一行，每一个变量对应矩阵中的一列，也可以利用 WHERE 语句为观测读取增加限定条件。

1. 打开与激活数据集

数据集在使用之前需进行打开操作，这里提供三种打开数据集的操作方法。

(1) 对于仅需要读入已存在数据集的情形，可以使用 USE 语句，其引用的一般形式为：

```
USE SAS-data-set <VAR operand> <WHERE (expression)>;
```

获得读权限时，可使用的语句包括 FIND、INDEX、LIST、READ 等。

(2) 对于需要读写已存在数据集的情形，可以使用 EDIT 语句，其引用的一般形式为：

```
EDIT SAS-data-set <VAR operand> <WHERE (expression)>;
```

获得读写权限时，可使用的语句包括前面的 FIND 等，以及 REPLACE、APPEND、DELETE、PURGE 等。

(3) 对于需要创建新数据集的情形，可以使用 CREATE 语句，其引用的一般形式为：

```
CREATE SAS-data-set <VAR operand>;
CREATE SAS-data-set FROM from-name
< [COLNAME = column-name ROWNAME = row-name]>;
```

可以使用 APPEND 语句将矩阵数据添加到数据集中。

IML 执行的指令都只对当前激活的数据集有效，所以连续对同一个数据集进行操作时无须每次都进行指定。当前激活的数据集通常有两个：一个用于输入，另一个用于输出。当一个数据集被打开时，IML 会默认它为当前激活数据集。下面几种途径用于激活数据集：

- (1) USE 和 SETIN 语句激活一个数据集作为当前输入。
- (2) SETOUT 语句激活一个数据集作为当前输出。
- (3) CREATE 和 EDIT 语句激活一个数据集作为当前输入和输出。

2. 显示与引用数据集

可以使用 SHOW 语句来显示一些关于数据集的有用信息，SHOW DATASETS 语句可以列出所有已打开数据集和它们的状态，SHOW CONTENTS 语句可以列出当前输入数据集中的观测数、变量数以及各变量名称、类型、大小等信息。

上一节对数据集的各种操作运算对象都是数据集名称，这个名称可以是一级的，也可以是二级的。对于二级名称，第一级应是 SAS 数据库名，第二级应是 SAS 数据集名，WORK 库是一个临时数据库，SAS 关闭后将被清空。

如果只使用一级名称，IML 将会指定一个默认的数据库。在 IML 运行初始阶段，默认数据库通常是 WORK，可以利用 RESET DEFLIB 语句重置默认数据库。如果试图创建一个永久数据集，就必须使用二级名称并且重置默认的数据库。

【例 3-157】 数据集的打开、激活、显示与引用。设程序名为 SASTJFX3_157.sas。

```
data a1;
  input id name $ score@@ ;
  cards;
  1 ZS 94 2 FQ
  3 NQ 99 4 DZ
  5 LR 95 6 NS
  7 SS 97 8 FS
  9 DK 98 10 ND
  ;
run;

proc iml;
  use A1;
  show datasets;
  edit A2;
  show datasets;
  setin A1 point 5;
  show datasets;
  show contents;
run;quit;
```

运行结果：

| | | | |
|---------|---------|-----------|----------------------|
| LIBNAME | MEMNAME | OPEN MODE | STATUS |
| WORK | A1 | Input | Current Input |
| LIBNAME | MEMNAME | OPEN MODE | STATUS |
| WORK | A1 | Input | |
| WORK | A2 | Update | Current Input/Output |
| LIBNAME | MEMNAME | OPEN MODE | STATUS |
| WORK | A1 | Input | Current Input |
| WORK | A2 | Update | Current Output |

DATASET: WORK. A1. DATA

| | | |
|----------|------|------|
| VARIABLE | TYPE | SIZE |
| id | num | 8 |
| name | char | 8 |
| score | num | 8 |

Number of Variables : 3
Number of Observations: 10

3. 选择观测条目

数据集中的变量和观测可以通过 LIST 语句列出，其引用的一般形式为：

```
LIST <range> <VAR operand> <WHERE (expression)> ;
```

(1) 列出指定范围观测

可以通过一些关键字或者 POINT 选项输出指定的观测条目，参数 range 常用的关键字有以下几个：

- ①ALL 指定所有观测。
 - ②CURRENT 指定当前观测。
 - ③NEXT <number> 指定下一条或几条观测。
 - ④AFTER 指定当前观测后面的所有观测。
- POINT operand 指定被选择的某条观测，参

表 3-30 operand 常见举例

| operand | 举例 |
|-----------|---------------|
| 单个记录号 | point 7 |
| 一系列记录号 | point {3 5 7} |
| 包含记录号的矩阵 | point matrix |
| 表示记录号的表达式 | point (p + 1) |

见表 3-30。

【例 3-158】 利用 list 语句列出指定观测。设程序名为 SASTJFX3_158.sas。

```
proc iml;
  use A1;
  m = {1,3,5};
  list all;list point 4;
  list next 2;list point m;
  list point (m+1);
run;quit;
```

运行结果：

| | | | | | |
|-----|------------|---------|-----|-----------|---------|
| OBS | id name | score | 4 | 4.0000 DZ | 96.0000 |
| 1 | 1.0000 ZS | 94.0000 | OBS | id name | score |
| 2 | 2.0000 FQ | 98.0000 | 5 | 5.0000 LR | 95.0000 |
| 3 | 3.0000 NQ | 99.0000 | 6 | 6.0000 NS | 99.0000 |
| 4 | 4.0000 DZ | 96.0000 | OBS | id name | score |
| 5 | 5.0000 LR | 95.0000 | 1 | 1.0000 ZS | 94.0000 |
| 6 | 6.0000 NS | 99.0000 | 3 | 3.0000 NQ | 99.0000 |
| 7 | 7.0000 SS | 97.0000 | 5 | 5.0000 LR | 95.0000 |
| 8 | 8.0000 FS | 98.0000 | OBS | id name | score |
| 9 | 9.0000 DK | 98.0000 | 2 | 2.0000 FQ | 98.0000 |
| 10 | 10.0000 ND | 97.0000 | 4 | 4.0000 DZ | 96.0000 |
| OBS | id name | score | 6 | 6.0000 NS | 99.0000 |

“point 4”使得第四条观测为当前观测，“next 2”自然输出随后两条观测，m 是一个矩阵，含有三个元素，list 语句输出对应的三条观测，m + 1 即所有元素都加 1，输出的三条观测也随之改变。

(2)选择变量

可以使用 VAR 语句来选择变量，其引用的一般形式为：

```
VAR operand
```

operand 可以是下列内容之一：

- ①变量名组成的集合。
- ②包含变量名的矩阵。
- ③表示变量名的表达式。
- ④特定关键字：_ALL_(所有变量)，_CHAR_(所有字符变量)，_NUM_(所有数值变量)。

【例 3-159】 利用 list 语句结合变量名选择观测。设程序名为 SASTJFX3_159.sas。

```
proc iml;
  use A1;
  m = {1,3,5};n = {name score};
```

```
list all var _all_;list point 4 var{id,score};
list next2 var _char_;list point m var n;
run;quit;
```

运行结果：

| | | | | | | | |
|-----|---------|----|---------|----------|---------|---------|---------|
| OBS | id name | | score | 10 | 10.0000 | ND | 97.0000 |
| 1 | 1.0000 | ZS | 94.0000 | OBS | id | score | |
| 2 | 2.0000 | FQ | 98.0000 | 4 | 4.0000 | 96.0000 | |
| 3 | 3.0000 | NQ | 99.0000 | OBS name | | | |
| 4 | 4.0000 | DZ | 96.0000 | 5 | LR | | |
| 5 | 5.0000 | LR | 95.0000 | 6 | NS | | |
| 6 | 6.0000 | NS | 99.0000 | OBS name | | score | |
| 7 | 7.0000 | SS | 97.0000 | 1 | ZS | 94.0000 | |
| 8 | 8.0000 | FS | 98.0000 | 3 | NQ | 99.0000 | |
| 9 | 9.0000 | DK | 98.0000 | 5 | LR | 95.0000 | |

_ALL_关键字提示输出所有变量，_CHAR_提示输出所有字符变量，例中只有 NAME 变量是字符的，矩阵 n 只包括两个元素，因此 IML 只输出与变量 NAME、SCORE 对应的数据信息。

(3)选择观测

利用 WHERE 子句可以有条件地选择观测，其引用的一般形式为：

```
WHERE variable comparison-op operand ;
```

有关的算符说明参见表 3-31。

表 3-31 comparison-op 常用算符说明

| 算符 | 功能 | 子句成立条件 | 算符 | 功能 | 子句成立条件 |
|-----|------|--------|-----|---------------|--------|
| < | 小于 | 所有元素 | ^= | 不等于 | 所有元素 |
| < = | 小于等于 | 所有元素 | ? | 包含给定字符串 | 任一元素 |
| > | 大于 | 所有元素 | ^? | 不包含给定字符串 | 所有元素 |
| >= | 大于等于 | 所有元素 | = : | 以给定字符串开始 | 任一元素 |
| = | 等于 | 任一元素 | = * | 与给定字符串发音或拼写类似 | 任一元素 |

注：子句成立条件针对的是以 operand 为矩阵的情形。
operand 可以是集合、矩阵或者表达式。
逻辑表达式可以用表示交集符号“&”和表示并集符号“|”来实现。

【例 3-160】 利用条件选择观测。设程序名为 SASTJFX3_160.sas。

```
proc iml;
use A1;m={94,98};n={name score};
list all var _all_ where(score={95 97});
list all var n where(score>m);
run;quit;
```

运行结果：

| | | |
|-----|-----------|---------|
| OBS | id name | score |
| 5 | 5.0000 LR | 95.0000 |
| 7 | 7.0000 SS | 97.0000 |

| | | |
|-----|------------|---------|
| 10 | 10.0000 ND | 97.0000 |
| OBS | name | score |
| 3 | NQ | 99.0000 |
| 6 | NS | 99.0000 |

从例子可以看出“>”与“=”在与矩阵比较时成立的条件是不相同的。这里要注意的是，WHERE 子句中表达式左边是来自于数据集的变量，表达式右边可以是标量、向量或矩阵，在一个表达式中使用的数据集变量不能超过一个，且只能用在表达式左侧。

4. 从数据集中读取观测

READ 语句可以实现观测从数据集到矩阵的传递，首先要确保待读取的数据集必须已经打开并且是当前被激活的输入数据集。打开数据集可以用 USE、EDIT 等语句，激活一个输入数据集可以用 SETIN 语句。

```
READ <range> <VAR operand> <WHERE (expression)> <INTO name>;
```

各参数意义与前面介绍的基本相同。

只使用 READ 语句和 VAR 子句从当前激活的输入数据集中读取变量时，VAR 子句中的每个变量都会得到一列向量，向量名称与对应的变量相同，列向量行数由选取的观测范围决定。

如果在此基础上添加 INTO 子句，就可以把 VAR 子句中的变量全部读入到一个指定的矩阵中。此时，每个变量都作为结果矩阵中的一列，矩阵行数取决于选取的观测数目。在观测选取上，可以结合 WHERE 子句有条件地进行选择。

【例 3-161】 从数据集有条件地读取观测到矩阵。设程序名为 SASTJFX3_161.sas。

```
proc iml;
  use A1;m={98,97};n={id score};
  reset print;
  read all var _char_where(score={96 95}) into test1;
  read all var n where(score > (m-1)) into test2;
run;quit;
```

运行结果：

| | | | | | | | |
|-------|--------|-------|---------------------|-------|--------|--------|-----------|
| test1 | 2 rows | 1 col | (character, size 8) | test2 | 5 rows | 2 cols | (numeric) |
| | | DZ | | | 2 | 98 | |
| | | LR | | | 3 | 99 | |
| | | | | | 6 | 99 | |
| | | | | | 8 | 98 | |
| | | | | | 9 | 98 | |

例中创建了两个矩阵：第一个矩阵包含数据集中所有字符变量，并且只读取 score 值为 96 或 95 的观测；第二个矩阵只包含数据集中变量名为 id、score 的两个变量，并且只读取 score 值大于 97 的观测。

5. 编辑 SAS 数据集

通常使用 EDIT 语句获得对一个数据集的读/写操作，常见的编辑操作包括更新变量值、标记待删除观测、删除已标记观测、保存修改等。

假设要更新某个数据集中的变量值，应遵循以下的步骤：

- (1) 获得待编辑数据集的读写权限。
- (2) 寻找待修改的观测。
- (3) 修改指定的变量值。
- (4) 替代数据集中的原始值。

【例 3-162】 修改指定观测的变量值。设程序名为 SASTJFX3_162. sas。

```
proc iml;
edit A1; reset print;
find all where(score = {95 96}) into test1;
list point test1;
score=90;replace point test1;
list point test1;
run;quit;
```

运行结果：

| OBS | id name | score | OBS | id name | score |
|-----|-----------|---------|-----|-----------|---------|
| 4 | 4.0000 DZ | 96.0000 | 4 | 4.0000 DZ | 90.0000 |
| 5 | 5.0000 LR | 95.0000 | 5 | 5.0000 LR | 90.0000 |

首先，利用 find 语句找到满足特定条件的观测，最终存到矩阵 test1 是一切满足条件的观测标号；然后，修改 x 的值为 500，利用 replace 语句替代 test1 中对应观测的 x 值，即完成了数据集的更新。

有时，可能需要删除某些指定的观测，这时要用到 DELETE 语句，其引用的一般形式为：

```
DELETE <range> <WHERE(expression)>;
```

常见几种 DELETE 语句的用法如下。

- (1) Delete: 删除当前观测。
- (2) Delete point 3: 删除第 3 条观测。
- (3) Delete all where(x > 3); 删除所有 x 大于 3 的观测。

如果一个数据集中已经删除了很多观测，直接导致余下的观测标号不连续，不利于后续操作，则可以使用 PURGE 语句对观测重新进行编号。

【例 3-163】 删除指定观测。设程序名为 SASTJFX3_163. sas。

```
proc iml;
edit A1;
find all where(score >= 94 & score <= 98) into test1;
list point test1;
delete point test1;list all;
purge;list all;
run;quit;
```

运行结果：

| OBS | id name | score | OBS | id name | score |
|-----|-----------|---------|-----|------------|---------|
| 1 | 1.0000 ZS | 94.0000 | 6 | 5.0000 LR | 97.0000 |
| 2 | 2.0000 FQ | 98.0000 | 7 | 6.0000 NS | 98.0000 |
| 4 | 3.0000 NQ | 96.0000 | 8 | 7.0000 SS | 98.0000 |
| 5 | 4.0000 DZ | 95.0000 | 9 | 8.0000 FS | 97.0000 |
| | | | 10 | 10.0000 ND | 97.0000 |

| OBS | id name | score | OBS | id name | score |
|-----|-----------|---------|-----|-----------|---------|
| 3 | 3.0000 NQ | 99.0000 | 1 | 3.0000 NQ | 99.0000 |
| 6 | 6.0000 NS | 99.0000 | 2 | 6.0000 NS | 99.0000 |

首先利用 FIND 语句找到了满足条件的观测标号，然后将其删除。可以发现，由于删除操作使得 OBS 一列不连续，这不利于后续的更新或者删除观测等操作的进行，因此用 PURGE 语句对余下观测重新进行排号。

6. 由矩阵创建数据集

IML 允许用户由矩阵创建数据集，应用的语句主要有 CREATE 和 APPEND，原矩阵的列将变成新数据集的变量，原矩阵的行将变成新数据集的观测，即 $m \times n$ 矩阵将得到一个 m 行观测 n 个变量的数据集。CREATE 语句创建了一个新数据集并且获得对它的读/写权限，APPEND 语句将矩阵内容写入到观测中。

```
CREATE SAS-data-set FROM matrix < [COLNAME = column-name ROWNAME = row-name] >;
```

COLNAME 为数据集中变量命名，ROWNAME 增加一个包含行标题的变量。

【例 3-164】 利用 CREATE 语句由矩阵创建数据集。设程序名为 SASTJFX3_164.sas。

```
proc iml;
  use A1;reset print;
  read all var _NUM_ into test;
  show datasets;
  create A2 from test[colname={'id' 'score'}];
  append from test;
  show datasets;show contents;
run;quit;
```

运行结果：

| | | | |
|----------------------------|---------|-----------|----------------------|
| LIBNAME | MEMNAME | OPEN MODE | STATUS |
| WORK | A1 | Input | Current Input |
| LIBNAME | MEMNAME | OPEN MODE | STATUS |
| WORK | A1 | Input | |
| WORK | A2 | Update | Current Input/Output |
| DATASET : WORK. A2. DATA | | | |
| VARIABLE | TYPE | SIZE | |
| id | num | 8 | |
| score | num | 8 | |
| Number of Variables : 2 | | | |
| Number of Observations: 10 | | | |

从结果可以看出，数据集 A2 由 CREATE 创建并且成为当前激活的读/写数据集，但此时 A2 中并没有观测，需要 APPEND 语句将观测写入到数据集中，最终数据集 A2 中含有 2 个变量和 10 条观测。

上面介绍的是 CREATE 语句和 FROM 子句的组合，实际应用中更为常见的是 CREATE 语句和 VAR 子句的组合。相比之下，后者具有更强的通用性。当要读取的变量既有数值型又有字符型时，只能用 VAR 子句。上例程序 FROM 子句部分可以用 VAR 子句修改如下：

```
create A2 var{id score};
append from test;
```

7. 数据集排序

IML 可以根据用户指定的关键变量对观测进行排序。对数据集排序前，要确认这个数据集是否被打开，只有关闭的数据集才能进行排序。SORT 语句和 by 子句再加指定变量就可以根据用户需求对数据集进行排序。如果要求保留原始数据，只需在排序的同时定义一个输出数据集即可。指定的排序变量可以是多个，关键字 DESCENDING 可以实现降幂排序。

【例 3-165】 用 SORT 语句对数据集进行排序。设程序名为 SASTJFX3_165. sas。

```
proc iml;
  show datasets;
  use A1;list all;
  close A1;
  sort A1 by score; /* sort A1 out =AA by score;* /
  use A1;list all;
  run;quit;
```

运行结果：

| OBS | id name | score | OBS | id name | score |
|-----|------------|---------|-----|------------|---------|
| 1 | 1.0000 ZS | 94.0000 | 1 | 1.0000 ZS | 94.0000 |
| 2 | 2.0000 FQ | 98.0000 | 2 | 5.0000 LR | 95.0000 |
| 3 | 3.0000 NQ | 99.0000 | 3 | 4.0000 DZ | 96.0000 |
| 4 | 4.0000 DZ | 96.0000 | 4 | 7.0000 SS | 97.0000 |
| 5 | 5.0000 LR | 95.0000 | 5 | 10.0000 ND | 97.0000 |
| 6 | 6.0000 NS | 99.0000 | 6 | 2.0000 FQ | 98.0000 |
| 7 | 7.0000 SS | 97.0000 | 7 | 8.0000 FS | 98.0000 |
| 8 | 8.0000 FS | 98.0000 | 8 | 9.0000 DK | 98.0000 |
| 9 | 9.0000 DK | 98.0000 | 9 | 3.0000 NQ | 99.0000 |
| 10 | 10.0000 ND | 97.0000 | 10 | 6.0000 NS | 99.0000 |

8. 建立数据集索引

对于规模较大的数据集，要查询某个指定变量的信息通常要花费较长时间，因为 SAS 要逐条读取观测，我们可以通过事先建立索引变量的方式来减少查询时间。INDEX 语句将建立一个包含变量值和索引变量记录号的特殊文件，IML 应用这个索引可以进行更高效的查询。我们可以为任意变量建立索引，但一次只能使用一个索引，而且在应用 PURGE 语句过后会丢失所有索引关联，IML 会自动删除全部索引。

索引变量值会自动更新，因此不用担心每次修改数据集之后都要重新建立一次索引。当使用索引时，观测顺序通常已经被打乱，因此不能根据原始物理标号对观测随意访问，也就是说 POINT 语句在执行索引时是禁用的。

【例 3-166】 为数据集变量建立索引。设程序名为 SASTJFX3_166. sas。

```
proc iml;
  use A1;list all;
  index name;
  list all;
  run;quit;
```

运行结果：

| OBS | id name | score | OBS | id name | score |
|-----|------------|---------|-----|------------|---------|
| 1 | 1.0000 ZS | 94.0000 | 9 | 9.0000 DK | 98.0000 |
| 2 | 2.0000 FQ | 98.0000 | 4 | 4.0000 DZ | 96.0000 |
| 3 | 3.0000 NQ | 99.0000 | 2 | 2.0000 FQ | 98.0000 |
| 4 | 4.0000 DZ | 96.0000 | 8 | 8.0000 FS | 98.0000 |
| 5 | 5.0000 LR | 95.0000 | 5 | 5.0000 LR | 95.0000 |
| 6 | 6.0000 NS | 99.0000 | 10 | 10.0000 ND | 97.0000 |
| 7 | 7.0000 SS | 97.0000 | 3 | 3.0000 NQ | 99.0000 |
| 8 | 8.0000 FS | 98.0000 | 6 | 6.0000 NS | 99.0000 |
| 9 | 9.0000 DK | 98.0000 | 7 | 7.0000 SS | 97.0000 |
| 10 | 10.0000 ND | 97.0000 | 1 | 1.0000 ZS | 94.0000 |

如果对变量 score 建立索引，得到结果与按变量 score 对数据集排序得到的结果很类似。二者的区别是，IML 建立索引时会自动生成一个文件，不覆盖原数据集，对数据集排序时如果不额外定义输出数据集，那么原数据集将被覆盖；索引得到的特殊文件其观测的物理标号 (OBS) 并没有发生改变，仍与原文件相同，而排序之后的数据集物理标号已经重新编排。

9. IML 数据集操作与 DATA 步的比较

如果要在 IML 环境下模仿 DATA 步的处理方式，则必须明确 IML 与 SAS DATA 步之间的基本区别。

(1) 在 IML 环境下，通常以 CREATE 语句作为开头，而不是 DATA 语句。在建立数据集之前，必须明确所有待建立变量的属性，字符变量要预设好足够的字符串长度上限，IML 默认任意没有被声明字符型的变量都是数值型的；在 DATA 步中，变量属性是可以通过上下文来定义的。

(2) 在 IML 环境下，必须使用 APPEND 语句来输出观测；在 DATA 步中，或者使用 OUTPUT 语句来输出，或者随 DATA 步自动输出。

(3) 在 IML 环境下，需要使用 DO DATA 来实现循环；而 DATA 步中，循环是默认执行的。

(4) 在 IML 环境中，可以使用 CLOSE 语句执行数据集关闭操作，或者直接用 QUIT 语句退出 IML 环境；DATA 步会在循环结束后自动关闭数据集。

(5) DATA 步执行速度通常比 IML 要快。

简而言之，DATA 步处理问题更简易，需要的程序更简洁，但 IML 是交互式的，其强有力的矩阵处理能力大大增加了它在应用中的灵活性。数据管理指令见表 3-32。

表 3-32 数据管理指令

| 指 令 | 功 能 | 指 令 | 功 能 |
|--------|----------------------|---------------|---------------------|
| APPEND | 在数据集最后添加观测 | REPLACE | 替换数据集中的观测 |
| CLOSE | 关闭数据集 | RESET DEFLIB | 指定默认 SAS 库 |
| CREATE | 创建并打开一个数据集作为读/写数据集 | SAVE | 保存更改并重新打开数据集 |
| DELETE | 标记数据集中删除的观测 | SETIN | 激活一个已经打开的数据集作为输入数据集 |
| EDIT | 打开一个已经存在的数据集作为读/写数据集 | SETOUT | 激活一个已经打开的数据集作为输出数据集 |
| FIND | 搜索观测 | SHOW CONTENTS | 显示当前输入数据集的目录 |
| INDEX | 为数据集变量建立索引 | SHOW DATASETS | 显示当前打开的数据集 |
| LIST | 显示观测 | SORT | 对数据集排序 |
| PURGE | 清除数据集中所有已删除的观测 | SUMMARY | 产生数值变量的概括统计量 |
| READ | 将观测读入 IML 变量 | USE | 打开一个已经存在的数据集作为输入数据集 |

(胡纯严 郭辰仪 吕辰龙 曹波 郭春雪)

第2篇 统计设计中关键技术的 SAS 实现

第4章 统计设计核心内容介绍

从医学统计学本身来说，其包含内容众多，有统计设计、质量控制、设计类型、样本含量估计、表达描述、统计分析、结果解释及结论陈述等。本章简要介绍统计设计中的核心内容。

4.1 统计设计概述

科研设计是为了尽可能圆满地完成事先确定的研究目的一项科学研究的一切考虑和安排。科研设计包括专业设计与统计设计。统计设计由“试验设计、临床试验设计和调查设计”三种类型的设计组成。本节将简要介绍统计设计有关的内容。

4.1.1 统计设计类型

统计设计由“试验设计、临床试验设计和调查设计”三种类型的设计组成。

4.1.2 三类统计设计的共性

试验设计、临床试验设计和调查设计之间有很多共同的内容。

(1) 目标相同。都是为了更好地完成事先确定的研究目标而从统计学角度所做的考虑、计划和安排，都希望能透过事物的现象看清其本质。

(2) 核心内容相同。一般都不可避免地涉及“三要素”、“四原则”和“设计类型”。

(3) 应有严格质量控制。在科研工作具体实施之前，都应对未来的试验或调查过程中可能出现的意想不到的困难、干扰和影响结果正确性的混杂因素，有足够的警觉和必要的防范措施（可简称为过程中的质量控制），以确保研究方案的顺利实施和研究结果的正确性与稳定性。

4.1.3 三类统计设计的个性

由于研究的场合、对象和方法等不完全相同，试验设计、临床试验设计和调查设计又具有各自的特性，称其为“个性”。此处，将它们各自的个性分述如下。

(1) 试验设计：一般指在试验室进行的小规模试验研究所对应的设计。其受试对象通常是动物或样品，各试验因素可由研究者根据基本常识和专业知识选定，非试验因素也比较容易控制。因此，若采用恰当的试验设计类型，可以同时考察很多试验因素及其交互作用对观测结果的影响大小，可以将重要的非试验因素的干扰和影响控制在最低水平，过程中的质量可以控制

得很严,组间的可比性好,数据的说服力强,所获得的结果准确度高,重现性好。一般来说,所得结论更能经得起时间与实践的检验。

(2) 临床试验设计:一般来说,其受试对象是正常人或病人。研究者必然要面对“伦理道德和受试者依从性”两大难题的挑战。怎样设置合理的对照组、如何严格遵守伦理道德、如何选择和剔除受试者、如何提高受试者的依从性、如何控制人为因素(特别是心理因素)引起的偏性等问题的合理解决,就是临床试验设计的最重要的内容。

(3) 调查设计:在没有任何人为干预的前提下,对客观存在的事物或现象或受试对象进行被动的观察,不可避免地涉及调查对象的地域、时空分布和对于总体的代表性问题。要根据拟完成的调查任务,结合基本常识、专业知识和统计学知识,把现场可能碰到的各种问题尽可能考虑周到,以免在调查结束后,面对“漏项和缺项无法弥补”所带来的尴尬,甚至导致调查研究前功尽弃的悲惨结局。对于一些敏感性调查问题,应采取切实可行的技术方法,以提高调查结果的可信度。调查表内容的设置是至关重要的,既不能过于简单或笼统(影响调查质量),又不能过于复杂或细致(影响调查表的回收率)。

4.1.4 试验设计要点

通常,在试验室内完成的以动物或样品为受试对象的研究称为试验研究。试验设计方案中的要点包括试验设计三要素、四原则、设计类型和质量控制。

1. 试验设计的三要素

在统计学中,常把“受试对象、影响因素和试验效应”称为试验设计的三要素。

(1) 受试对象:试验因素的承受者被称为受试对象。受试对象可以是人、动物或植物,也可以是某个器官、组织、细胞、亚细胞或血清等生物材料。无论受试对象是动物还是人,首先应满足以下三个条件:①对试验因素敏感;②反应必须稳定,如研究某药物对高血压的治疗效果,常选用 I、II 期高血压患者作为受试对象,因为 III 期高血压患者对药物不够敏感;③受试对象具有同质性和代表性。此外,不同的试验,还有一些特殊的纳入和排除标准。受试对象的数量通常指试验研究中总共需要多少样本含量,也称样本大小,在统计学上称为“样本含量估计”问题,可详见本章 4.4 节,此处从略。

(2) 影响因素:在试验中,研究者希望着重考察的某些试验条件和某些重要的非试验条件被称为影响因素。影响因素包括试验因素和非试验因素。试验因素是研究者希望通过试验着重考察的,而那些与试验因素同时存在、能使受试对象产生效应的其他因素被称为非试验因素,一般是受试对象自身具有的、可影响其发展过程的某些属性(如性别、年龄、血型等)或状态(如病型、病情、患病时间、生理或心理所处的时期等)。因考察的效应指标不同,有些非试验因素对试验结果的影响是不容忽视的(被称为重要的非试验因素),必须在试验设计阶段或在统计分析阶段加以控制;而有些非试验因素是可以忽略不计的(被称为次要的非试验因素)。

(3) 试验效应:试验效应就是试验因素作用于受试对象后产生的效果,它是通过试验中所选用的指标来体现的。所选用的指标与要反映的问题之间应具有较高的关联性,判断指标取值大小时应具有较高的客观性、特异性、灵敏性和精确性。因此,在选用指标时,应尽量多选定定量指标或少数量化起来较为方便的定性指标。

2. 试验设计的四原则

随机、对照、重复和均衡是试验设计的四个基本原则。

(1) 随机原则: 所谓随机原则就是在抽样或分组时必须做到使总体中任何一个个体都有同等的机会被抽取进入样本, 以及样本中任何一个个体都有同等机会被分配到任何一个组中。在受试对象的选取和分组时, 必须严格按这一原则实施。随机化分为简单随机化、系统随机化、分层随机化、分层区组随机化等。实现随机化的方法有多种, 如抽签、查随机数字表或随机排列表、利用计算机产生伪随机数。通过随机化, 降低系统误差的影响。现在, 随机化原则的具体实施一般都是通过统计软件来实现的。与随机化有关的内容将在本书中第7章中专门介绍。

(2) 对照原则: 进行试验研究, 必须设立对照组。因为有比较才有鉴别, 缺少对照的研究是没有说服力的。通过对照来鉴别处理因素与非处理因素之间的差异, 抵消或减少试验误差。对照原则讲起来容易, 在实践中做的难度是比较大的。因此, 实际工作者经常在对照原则方面出现这样或那样的问题, 这一点务必要引起读者的高度重视。

(3) 重复原则: 由于个体差异等影响因素的存在, 同一种处理对不同的受试对象所产生的效果不尽相同, 其具体指标的取值必然有高低之分, 只有在大量重复试验的条件下, 该处理的真实效应才会比较确定地显露出来。因此, 在试验研究中, 必须坚持重复的原则。

“重复”一词在试验研究中至少有下面三层含意: 其一, 对某样品进行观测时, 为了减少方法和操作等带来的误差, 将每一个样品分成 k 份, 测出各份样品单位容积中某指标的观测值, 用它们的算术均数作为每份样品单位容积中该指标的直接观测值; 其二, 在相同的试验条件下, 独立重复地观测 m 个样品(或受试对象), 这就是人们通常所指的“重复试验”, 其目的是为了降低以个体差异为主的各种试验误差; 其三, 在部分或全部试验条件有规律地变动时, 从同一个样品(或受试对象)上重复测量到 k 个数值, 称为具有重复测量的试验或设计。这种重复最有利于显示指标的动态变化情况。整个试验过程中不同受试对象的个数称为样本含量或样本大小。样本含量 n 过大(有时人力不够, 工作粗枝大叶, 资料可靠度低)或过小(统计规律无法显露出来)都有弊病, 最好针对具体情况, 根据专业知识和统计学知识做出合理的估计。重复原则最重要的一个具体实施就是“样本含量估计”问题, 可详见本章 4.4 节, 此处从略。

(4) 均衡原则: 所谓均衡, 就是在单因素试验研究中, 设法使对照组与试验组中的非试验因素尽量达到均衡一致, 使试验因素的试验效应能更真实地反映出来; 而在多因素试验研究场合下, 对照是指同一个试验因素各水平之间互为对照。可以说, 均衡性原则是试验设计中最易被人们忽视的, 同时又是最重要的原则。很多统计研究设计方案, 若在均衡性方面考虑不周到, 其研究结论就很容易被推翻。

3. 试验设计的设计类型

试验设计类型就是因素及其水平所决定的一种结构。这种结构不单纯取决于“形式”, 还取决于“内容”。具体见本章 4.2 节, 此处从略。

4. 试验设计的质量控制

对试验过程中出现的可能会影响结果准确度的任何问题, 都应能及时发现, 并能在尽可能短的时间内消除其影响。需要进行质量控制的环节包括: 控制研究者的责任心、心理状态和技术水平对试验结果可能造成的不利影响; 控制受试者的心理状态和不依从性对试验结果可能造成的不利影响; 控制环境的变化(如温度、湿度、光线等)对试验结果可能造成的不利影响; 控制试验条件的改变(如仪器的状态发生改变、试剂过期、观测指标的单位不同、测定的方法不同等)对试验结果可能造成的不利影响。

对试验过程中的质量进行控制的措施有: 对不同单位的多名研究者应进行统一的培训, 内容

包括增强责任心教育、调适好心理状态(在言行上不应有个人的倾向性或偏好)、接受正规的技术培训,并要求每一位研究者自觉遵守统一的操作规程。对于动物或样品类的受试对象而言,应当选用与研究目的和试验因素相对应(同质)的动物或样品;对于某病患者的受试对象而言,应当制定严格的纳入与排除标准;应尽可能采取双盲或三盲法分配受试对象,尽可能减少或消除来自研究者,特别是受试者心理的影响。对于环境因素的影响,应尽可能创造条件,使试验研究在特定的温度、湿度和光线条件下进行;对于仪器设备、试剂应尽可能校准,并经常对其进行检查和调试;对于指标的测定单位、测定方法应完全统一,实在不能统一的,应采取合适的方法进行换算;对于主要评价指标,若测定方法无法统一,应确定一个有资质的测定部门来完成测试,称为“中心试验室”,这是消除多个试验研究单位测定结果参差不齐的重要措施之一。

4.1.5 临床试验设计要点

临床试验是指在人体(患者或健康志愿者)中进行的试验药物、医疗器械或治疗方法的系统性研究,通过比较试验组与对照组的试验结果,揭示试验药物、医疗器械或治疗方法对人体的作用、不良反应,或探索药物在人体内的吸收、分布、代谢和排泄规律等的一种前瞻性研究。临床试验与试验室内进行的试验明显的区别是受试对象不同。临床试验分为 I、II、III、IV 期。新药在批准上市前,通常应当进行 I、II、III 期临床试验;IV 期临床试验在新药上市后进行。

1. 临床试验的伦理道德问题

临床试验应遵守有关的法规和指南,如药品管理法、药品注册管理办法、新药审批办法、药品临床试验管理规范(GCP)等。同时,所有以人为受试对象的研究必须符合《赫尔辛基宣言》和国际医学科学组织委员会颁布的《人体生物医学研究国际道德指南》的道德原则。2008 年 10 月,《赫尔辛基宣言》修正案扩展了宣言的适用对象,重申并进一步澄清了基本原则和内容,加强了对受试者的权利保护,同时还增加了临床试验数据注册和使用人体组织时的同意等新内容,提高了人体医学研究的伦理标准。

在药物和医疗器械临床试验的过程中,必须对受试对象的个人权益给予充分的保障,并确保试验的科学性和可靠性。研究者必须向受试对象提供口头或书面的有关临床试验的详细材料,包括试验目的、预期收益、受试对象被分配到不同处理组而可能发生的风险与不便、因参加试验而受到损害或影响身体健康时能够获得的治疗和补偿。研究者不能强迫受试对象参加试验,经受试对象同意后,须由受试者或其法定代理人在知情同意书上签字并注明日期,执行知情同意过程的研究者或其代表也需在知情同意书上签名,并注明日期。知情同意书应使用受试对象能够懂得的语言和文字,受试对象保留在任何时候退出试验的权利。

2. 诊断标准、纳入标准与排除标准

(1)诊断标准:尽可能客观、量化,最好选择国际公认的标准。在临床试验过程中,不论是选择病例,还是判断效果,诊断标准一经采用,就必须自始至终地去遵照执行,不可随意改动。

(2)纳入标准:应该根据研究目的和实际情况制定。标准定得太高,增加工作量,不易找到研究对象。标准定得太低,又会影响研究结果。在制定纳入标准时,应尽可能地选择对干预措施有反应的病例作为研究对象。

(3)排除标准:临床试验中,通常有下列情况的患者不能作为研究对象。

①同时患另一种可能影响本试验效果的疾病的患者。

②同时患其他严重疾病者,因为这样的患者可在研究过程中死亡或因病情恶化而被迫退出。

③已知对药物有不良反应者。

④具有某些特殊情况者(已怀孕者、婴幼儿、具有可能降低某些依从性的某些残疾等)。

3. 受试对象的选择

临床病例选择一般应恪守三个一般原则:第一,不违反伦理道德;第二,制定出合理的标准;第三,具有普适性与可行性。要求:①入选的研究对象确能从科研中受益;②研究对象具有代表性并且依从性好。同时,志愿者选为研究对象的问题值得商榷。因为志愿者多有难言的苦衷,如经济困难或自身患病求医心切等。这些均可对研究结果的正确评价产生不利的影响,甚至可出现假阳性应答。再者,志愿者代表性差。

4. 主要疗效指标与安全性指标

(1)疗效指标:可分为主要疗效指标和次要疗效指标。主要疗效指标是与试验目的有本质联系的,能确切反映药物有效性或安全性的观察指标。通常主要指标只有一个。主要指标应根据试验目的选择易于量化、客观性强、重复性高,并在相关研究领域已有公认的标准指标。次要指标是指与试验目的相关的辅助性指标。

(2)安全性指标:在安全性指标较多时,也应指出主要与次要安全性指标分别是什么。药物安全性评价的常用统计指标为不良事件发生率和不良反应发生率。

5. 盲法

盲法(blinding)作为纠正偏倚(bias)的一个重要措施,特别在临床试验中,是为了避免研究者与受试者的主观和心理因素对试验结果的干扰。临床试验根据设盲的程度分为单盲、双盲、三盲和非盲(开放性试验)。仅为受试者不知所接受的是何种疗法时,称单盲(a single blinding);而参与临床试验和判断疗效的医护人员与受试者都不知任何一位受试者所接受的为何种疗法时称双盲(a double blinding),从而使研究结果得以客观评价;三盲(a triple blinding)则是研究者(包括医护人员)、研究对象和负责资料收集和分析的人员均不知任何一位受试者所接受的为何种疗法,从而更好地避免了偏倚。除开放性试验(open label or unblinding)和某些不宜设盲的试验如外科手术、引起生活方式改变等干预试验外,一般均应采用盲法。

6. 临床试验的四原则

随机、对照、重复和均衡是临床试验设计的四个基本原则。该四原则的遵循与试验设计基本相同,这里不赘述。所不同的是在临床试验中,样本含量确定的原则是:如果经过统计学计算,样本含量少于《药品注册管理办法》的规定,按照国家要求确定;否则,按照统计学计算结果确定。

7. 临床试验的设计类型

试验设计类型就是因素及其水平所决定的一种结构。这种结构不单纯取决于“形式”,还取决于“内容”。具体见本章4.2节,此处从略。

8. 多中心临床试验与中心效应

多中心临床试验系指由一个单位的主要研究者总负责,多个单位的研究者合作,按同一个试验方案同时进行的临床试验。多中心试验可以在较短的时间内入选所需的病例数,且入选的病例范围广,临床试验的结果更具代表性。但影响因素也随之更复杂。

多中心临床试验有时会出现中心效应(即不同中心的观测结果之间差异有统计学意义)。

因此,进行多中心试验必须在统一的组织领导下,遵循一个共同制定的试验方案完成整个试验。各中心试验组和对照组病例数的比例应与总样本的比例相同,以保证各中心齐同可比。多中心试验要求各中心的研究人员采用相同的试验方法,试验前对人员统一培训,试验过程要有监控措施,以保证他们的诊断、观测、评价上的一致性。当主要指标可能受主观影响时,需进行统一培训和一致性检验,在正式开展试验之前,应设法使之达到符合一致性的要求。当主要指标在各中心的实验室的检验结果有较大差异或参考值范围不同或度量单位不同时,应采取相应的措施,如统一由中心实验室检验、采取相同的参考值范围或相同的度量单位或推导出校正公式。在双盲多中心临床试验中,盲底是一次产生的,应按中心分层随机;当中心数较多且每个中心的病例数较少时,可统一进行随机,不按中心分层。

4.1.6 调查设计要点

调查性研究对受试对象不施加任何干预措施,是在完全“自然状态”下对研究对象的特征进行观察、记录,并对观察结果进行描述和对比分析。

1. 明确调查研究目的和调查指标

调查目的就是明确要解决哪些问题。调查指标要精选,重点要突出,尽量选用客观性强、灵敏度高、精确度好的定量指标。

2. 确定调查对象并估计样本含量

要根据调查目的和调查指标来确定调查对象,即划清调查总体的同质范围。明确的纳入标准和排除标准可以用来明确调查对象,同时也是控制选择偏倚的有效手段。

样本含量估计问题见本章 4.4 节。

3. 明确调查方法和调查方式

根据调查范围大小调查方法分为普查、抽样调查、典型调查,根据研究时间又可以分为横断面研究和纵向研究。按抽取样本的方式可分为概率抽样调查和非概率抽样调查。

获取资料的方式主要有直接观察法和采访法两种。

4. 拟定调查项目和调查表

根据调查指标确定调查项目,即调查的具体内容,包括分析项目和备查项目。

把调查项目按逻辑顺序列成表格形式供调查使用就是调查表。

5. 制订调查组织计划

调查的组织计划应包括组织领导、宣传动员、时间进度、分工协调、经费预算、运作管理、现场组织、调查员培训、调查表核查制度、资料汇总整理要求等。大规模的调查研究往往需要进行小范围的预调查,以便修改完善组织计划和调查设计方案。

6. 资料的整理和分析

资料的整理和分析包括审核复查、录入整理、统计分析三个过程。

7. 调查质量控制

调查研究中,调查结果常常出现误差,误差大致分为两类,即抽样误差和非抽样误差(包括系统误差和过失误差)。抽样误差在抽样调查中是不可避免的,有一定的规律,因此可以运用统计学方法进行估计。引起非抽样误差的原因较多,也比较复杂,控制起来较困难。对非抽样误差的控制应贯穿研究的整个阶段。

- (1)设计阶段：该阶段特别要注意明确研究范围和对象、选择恰当的调查指标、设计完善的调查表、采取合理的调查方式、开展必要的预调查等。
- (2)调查阶段：该阶段要求严格按照研究设计方案来执行。要重视调查人员的培训，提高其调查技巧和调研水平，最大限度地争取调查对象的配合，提高问卷的应答率，降低误答率。
- (3)资料整理和分析阶段：应采取有效的措施减少数据录入、汇总和计算等方面的错误，加强问卷审核、数据核查、结果复核等质量控制工作。

4.2 设计类型概述

试验设计类型，就是安排因素及其水平组合的一种结构。因素通常是指性质相同的不同试验条件的总称，而性质相同的不同试验条件又被称为该试验因素的不同水平。试验设计类型常需体现以下情况：因素及水平的数量、因素的性质、因素水平之间的组合关系、因素与受试对象之间的关系、因素是否同时施加、因素对结果的影响是否地位平等。正确识别试验设计类型，有利于正确选择统计分析方法处理实际资料。本节将介绍生物医学与临床研究中最为常用的几种试验设计类型。

4.2.1 单因素设计

试验中只存在一个试验因素的设计。可分为单组设计、配对设计和单因素 K 水平设计($K \geq 2$)。

1. 单组设计

比较某样本所代表的总体是否与已知总体相同。也就是比较样本所代表的总体的某平均值与已知总体该平均值之间的差别是否具有统计学意义。

【例 4-1】 已知玉米单交种群 105 的平均穗重为 300g。喷药后，随机抽取 9 个果穗，其穗重分别为：308, 305, 311, 298, 315, 300, 321, 294, 320g。问：喷药后与喷药前的果穗平均重量之间的差别是否具有统计学意义？

2. 配对设计

为控制非实验因素对结果的影响，将某些重要的非实验因素在实验组和对照组中进行匹配，可分为自身配对、同源配对和条件相近者配对三种类型。

【例 4-2】 对血小板活化模型大鼠以 ASA 进行试验性治疗，以血浆 TXB_2 (ng/L) 为指标，其结果见表 4-1。（此为自身配对）

【例 4-3】 从 8 窝仔猪中每窝选出性别相同、体重接近的两头仔猪进行饲料对比试验，将每窝两头仔猪随机分配到两个饲料组中，试验期 30 天，结果见表 4-2。（此为同源配对）

表 4-1 大鼠血小板活化模型 ASA 治疗前后血浆 TXB_2 的变化 (ng/L)

| 大鼠号 | 血浆 TXB_2 (ng/L) | |
|-----|-------------------|-----|
| | 给药前 | 给药后 |
| 1 | 250 | 184 |
| 2 | 226 | 205 |
| ⋮ | ⋮ | ⋮ |
| 10 | 176 | 176 |

表 4-2 喂饲不同饲料的仔猪体重 (kg)

| 窝号 | 仔猪体重 (kg) | |
|----|-----------|------|
| | 甲饲料 | 乙饲料 |
| 1 | 10.0 | 9.8 |
| 2 | 11.2 | 10.6 |
| ⋮ | ⋮ | ⋮ |
| 8 | 10.8 | 9.8 |

【例 4-4】 一位社会学者为了判定聪明人选择的配偶是否也聪明，随机选取 32 对夫妻，并测得智商得分(IQ)，见表 4-3。问：配偶之间差别是否有统计学意义？（此为条件相近者配对）

表 4-3 32 对夫妻的 IQ 得分

| 配偶号 | IQ 得分 | |
|-----|-------|-----|
| | 丈夫 | 妻子 |
| 1 | 95 | 95 |
| 2 | 103 | 98 |
| ⋮ | ⋮ | ⋮ |
| 32 | 96 | 102 |

3. 单因素 K 水平设计 ($K \geq 2$)

将受试对象随机分配成 K ($K \geq 2$) 组，每一组接受一种处理。K = 2 时，可称为成组设计。

【例 4-5】 一个小麦新品种经过 6 代选育，从第 5 代(A 组)中抽出 10 株，株高为：66、65、66、68、62、65、63、66、68、62(cm)，又从第 6 代(B 组)中抽出 10 株，株高为：64、61、57、65、65、63、62、63、64、60(cm)。问：株高性状是否已经达到稳定？（此为成组设计）

【例 4-6】 从津丰小麦 4 个品系中分别随机抽取 10 株，测量其株高(cm)，数据如下所示。问：不同品系津丰小麦的平均株高之间的差别是否具有统计学意义？

- 品系 0-3-1: 63、65、64、65、61、68、65、65、63、64
- 品系 0-3-2: 56、54、58、57、57、57、60、59、63、62
- 品系 0-3-3: 61、61、67、62、62、60、67、66、63、65
- 品系 0-3-4: 53、58、60、56、55、60、59、61、60、59

4.2.2 多因素设计

试验因素多于 2 个时的试验设计，按因素与水平的不同组合，可分为多种设计。

1. 随机区组设计

在试验中，涉及一个试验因素和一个区组因素(由一个或多个重要的非试验因素复合而成)，目的是减弱或消除非试验因素对结果的干扰和影响，更好地显露试验因素不同水平所产生的试验效应之间的差别。当每一区组内进入试验因素各水平组中的受试对象个数大于等于 2 时，为有重复试验的随机区组设计；当每一区组内进入试验因素各水平组中的受试对象个数为 1 时，为无重复试验的随机区组设计。

【例 4-7】 某研究者欲比较 3 种抗癌药物对小白鼠肉瘤的抑瘤疗效，先将 15 只染有肉瘤的小白鼠按体重大小配成 5 个区组，使每个区组内的 3 只小白鼠体重最接近，然后随机决定每个区组中的 3 只小白鼠接受 3 种药物的治疗，以肉瘤的重量为指标，试验结果见表 4-4。试比较不同抗癌药物对小白鼠肉瘤的抑瘤效果之间的差别是否有统计学意义。（此为无重复试验的随机区组设计）

表 4-4 不同药物作用后小白鼠肉瘤重量

| 区组 | 肉瘤重量(g) | | | |
|----|---------|------|------|------|
| | 药物种类: | A | B | C |
| 1 | | 0.82 | 0.65 | 0.51 |
| 2 | | 0.73 | 0.54 | 0.23 |
| 3 | | 0.43 | 0.34 | 0.28 |
| 4 | | 0.41 | 0.21 | 0.31 |
| 5 | | 0.68 | 0.43 | 0.24 |

【例 4-8】 现有三种降低转氨酶的药物 A、B、C，为了考察它们对甲型肝炎和乙型肝炎患者转氨酶降低的程度之间的差别是否有统计学意义，收集到的试验数据如表 4-5 所示(即从两型患者的总体中各随机抽取 9 例，然后，分别随机均分入三个药物组中去)，此为具有 3 次独立重复试验的随机区组设计(以前称为 3×2 析因设计，但不够严格，因为“型别”是一个属性因素)。其中，药物是试验因素，型别是属性因素。（此为有重复试验的随机区组设计）

表 4-5 三种药物对两型肝炎患者转氨酶降低效果的试验结果

| 肝炎类型 | | 转氨酶降低值 (kPa) | | | | | | | | |
|------|--|--------------|----|-----|-----|----|-----|----|-----|----|
| | | 药物种类: | | A 药 | | | B 药 | | C 药 | |
| 甲型 | | 100 | 85 | 90 | 120 | 90 | 110 | 50 | 60 | 40 |
| 乙型 | | 65 | 75 | 100 | 45 | 30 | 50 | 50 | 60 | 45 |

2. 平衡不完全随机区组设计

当试验因素的水平数大于每个区组内可以容纳的受试对象个数时，可选用平衡不完全随机区组设计，它是随机区组设计在解决实际问题时的一种变形或灵活运用。

【例 4-9】 用四种方法治疗脚气，受试者为两脚都患脚气的患者，每只脚接受一种处理，观察指标为治疗效果评分。设计格式和资料见表 4-6。

表 4-6 四种方法治疗脚气的评分

| 受试者 编 号 | 疗 效 评 分 | | | | |
|------------|---------|-----|-----|-----|-----|
| | 治疗方法： | 甲 | 乙 | 丙 | 丁 |
| 1 | | 5 | 3 | ... | ... |
| 2 | | ... | ... | 4 | 7 |
| 3 | | 2 | ... | 6 | ... |
| 4 | | ... | 4 | ... | 8 |
| 5 | | 3 | ... | ... | 7 |
| 6 | | ... | 2 | 5 | ... |

本例中， $v=4$ (药物有 4 种)、 $a=2$ (每位患者形成一个区组，每人只有两只脚)， $r=3$ (每种药物被重复使用的次数)， $b=6$ (区组个数，即本例中的患者人数)，满足 $rv=ab$ ，即 $3 \times 4=2 \times 6$ ；同时满足 $\lambda=r(a-1)/(v-1)=3(2-1)/(4-1)=1$ 为整数。故本例为平衡不完全区组设计。受试对象是 6 名患脚气病的患者，试验因素是“治疗方法”，区组因素为每位患者 (即个体差异)，定量的观测指标为“疗效评分”。

3. 无重复试验的双因素设计

两个试验因素各水平组合条件下只做一次试验的试验设计为无重复试验的双因素设计。

【例 4-10】 将落叶松苗木栽在四块不同的苗床(B)上，每块苗床的苗木又分别使用三种不同的肥料(A)以观察肥效差异，两年后于每一苗床的各施肥小区内用随机方法抽取一株，测其高度，资料见表 4-7。

表 4-7 不同肥料 A 和不同苗床 B 对落叶松苗高影响的测定结果

| 肥料 A | 苗高 (cm) | | | | |
|----------------|----------------------|----------------|----------------|----------------|-----|
| | 苗床 B: B ₁ | B ₂ | B ₃ | B ₄ | 合计 |
| A ₁ | 55 | 47 | 47 | 53 | 202 |
| A ₂ | 63 | 54 | 57 | 58 | 232 |
| A ₃ | 52 | 42 | 41 | 48 | 183 |
| 合计 | 170 | 143 | 145 | 159 | 617 |

4. 析因设计

具备下列 7 个特点的多因素试验设计可被称为析因设计。

(1) 试验因素的个数 ≥ 2 。

- (2) 每个试验因素的水平数 ≥ 2 且可以不等。
- (3) 不同的试验条件数(或组合数)等于全部试验因素的水平数之乘积。
- (4) 各试验条件下至少要做两次独立重复试验。
- (5) 全部受试对象被完全随机地分配进入任何一个试验条件组中去,各小组中的受试对象个数可以不等,但最好相等。
- (6) 试验时,任何一次试验都将涉及任何一个试验因素的某个水平,即全部试验因素同时施加。
- (7) 进行统计分析时,假定全部试验因素对观测结果的影响是地位平等的,即没有主次之分。

由上述特点,决定了析因设计分别具有一个明显的优点和缺点。优点:可以分析每个试验因素的主效应和试验因素之间各级交互作用的效应大小。缺点:总试验次数太多,试验费用增大且费时费力。

【例 4-11】 2×2 析因设计实例:将 40 只雌性小鼠完全随机地均分为 4 组,每组 10 只。第一组为空白对照组,即不给予任何药物治疗;第二组为仅给予一定剂量的激素治疗;第三组为仅给予一定剂量的某种中药治疗;第四组为既给予激素又给予中药(剂量同前面两组)治疗。观测指标为“排卵数”。设计格式见表 4-8。

表 4-8 中药用否与激素用否对雌性小鼠排卵数的影响情况

| 中药 用否 | 排卵数(个, $\bar{x} \pm s$) | |
|----------|--------------------------|------------------|
| | 激素用否: | 用 否 |
| 用 | 38.40 \pm 26.91 | 10.10 \pm 5.32 |
| 否 | 37.90 \pm 27.18 | 8.90 \pm 5.20 |

5. 含区组因素的析因设计

先将受试对象按重要非试验因素形成区组(每个区组中受试对象的个数等于试验因素的水平组合数),再将每个区组中的受试对象完全随机地均分入全部试验因素水平组合中去。

表 4-9 两种饲料的不同配方对猪平均日增重量的影响结果

| 随机区组 编号 | 平均日增重量($\times 500g$) | | | |
|------------|-------------------------|----------|------------|----------|
| | $A_1(B_1)$ | B_2 | $A_2(B_1)$ | B_2 |
| 1 | 1.38 | 1.52 | 1.22 | 1.11 |
| \vdots | \vdots | \vdots | \vdots | \vdots |
| 6 | 1.47 | 1.53 | 1.16 | 0.99 |

【例 4-12】 有人研究猪食用不同饲料对体重增加量的影响。A(大豆粉 + 不同含量的蛋白质): A_1 (加 14% 蛋白质)、 A_2 (加 12% 蛋白质); B(玉米中己氨酸的含量): B_1 (含 0.6% 己氨酸)、 B_2 (缺乏己氨酸),共有 4 种不同的饲料配方。用 24 头猪作为受试对象,按猪的初始体重由轻到重排序,并形成 6 个随机区组,每组中的

4 只猪体重接近,被随机地分入 4 个饲料组中去,在喂养期间,尽可能使每只猪每餐都能吃饱。设计格式和资料见表 4-9,该资料所取自的试验设计类型为含区组因素的析因设计。

6. 分式析因设计

在析因试验中,如果有理由假定某些高阶交互作用可被忽略,那么只要做同水平标准析因设计的全部试验点中的一部分,就可以获得主效应与低阶交互作用效应的估计值,这样的设计被称为分式析因设计,也称分数析因设计或析因设计的部分实施。

7. 嵌套设计

在一个涉及两个或两个以上试验因素时,若这些试验因素具备下列两种情形之一,则称此设计为嵌套(系统分组)设计:

- (1) 试验因素存在自然属性上的嵌套关系。
- (2) 试验因素对定量结果指标的影响存在主次关系。

8. 裂区设计

在一个涉及两个或两个以上试验因素时,若这些试验因素具备下列两种情形之一,则称此设计为裂区(分割)设计。

- (1) 试验因素施加于受试对象时存在先后顺序之分。
- (2) 整体试验由不同试验者在不同时间重复完成。
- (3) 整体试验在不同地域上重复完成。

(4) 有些试验因素的水平改变起来很困难(或很浪费),另一些试验因素的水平改变起来很方便(或很不浪费),则先安排前者。

9. 拉丁方设计

当实验中只考察一个 k 水平的实验因素,但同时又涉及两个都具有 k 个水平的区组因素,且它们之间的交互作用无统计学意义时所采用的一种实验设计类型叫拉丁方设计。先将实验因素的 k 个水平随机排列成 k 行 k 列的方阵,要确保实验因素的任何一个水平在任何一行和任何一列中仅出现一次;在方阵的 k 行的左边安排一个具有 k 水平的区组因素,并在其 k 列的上方安排另一个具有 k 水平的区组因素。

10. 交叉设计

当试验中涉及一个二水平的试验因素时,根据专业知识的需要,研究者希望该试验因素的两个水平先后作用于同一个受试对象,以便更好地考察该试验因素的效应,这时往往就需要用到交叉设计。交叉设计是生物医学科研中很常见的一类实验设计,它是一种特殊的自身对照设计,按预先设计好的实验顺序,在各个阶段对研究对象依次实施各种处理。

11. 正交设计

在安排多因素试验时,当任何两个试验因素的水平组合满足以下两个条件时,就称这样的设计为正交设计。第一,任何一个试验因素的不同水平出现的次数相同;第二,任何两个试验因素的水平组合所形成的数对出现的次数相同。 $L_8(2^7)$ 正交表见表 4-10。

表 4-10 $L_8(2^7)$ 正交表

| 试验号 | 列号: 1 | 2 | 3 | 4 | 5 | 6 | 7 | Y(试验结果) |
|-----|-------|---|---|---|---|---|---|---------|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | X |
| 2 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | X |
| 3 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | X |
| 4 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | X |
| 5 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | X |
| 6 | 2 | 1 | 2 | 2 | 1 | 2 | 1 | X |
| 7 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | X |
| 8 | 2 | 2 | 1 | 2 | 1 | 1 | 2 | X |

在表 4-10 中,任何一列“1”与“2”各出现 4 次;任何两列不同数对: (1, 1)、(1, 2)、(2, 1) 和 (2, 2) 各出现 2 次。满足上述两个条件的数阵,在数学上被称为满足正交性。

由此可知:所谓正交设计,实际上就是根据具体情况选择合适的正交表来安排多因素试验,特别是把注意力放在表头的安排上,简称为表头设计;并且将表中的“水平代码”转换成试验因素的“真实水平”。

正交表可以分为同水平的正交表和混合水平的正交表两大类。第一类为同水平的正交表，其中常用的有下列4种。

- (1) 二水平的正交表： $L_4(2^3)$ 、 $L_8(2^7)$ 、 $L_{16}(2^{15})$ 、 $L_{32}(2^{31})$ 、 $L_{64}(2^{63})$ 。
- (2) 三水平的正交表： $L_9(3^4)$ 、 $L_{27}(3^{13})$ 、 $L_{81}(3^{40})$ 。
- (3) 四水平的正交表： $L_{16}(4^5)$ 、 $L_{64}(4^{21})$ 。
- (4) 五水平的正交表： $L_{25}(5^6)$ 、 $L_{125}(5^{31})$ 。

第二类为混合水平的正交表，其中常用的有：

$L_8(4 \times 2^4)$ 、 $L_{12}(3 \times 2^4)$ 、 $L_{16}(4^2 \times 2^9)$ 、 $L_{16}(4^3 \times 2^6)$ 、 $L_{18}(2 \times 3^7)$ 、 $L_{18}(6 \times 3^6)$ 、 $L_{27}(9 \times 3^9)$
 $L_{32}(4^5 \times 2^{16})$ 、 $L_{36}(6 \times 3^{12})$ 、 $L_{32}(16 \times 2^{16})$ 、 $L_{24}(3 \times 4 \times 2^4)$ 、 $L_{36}(6^2 \times 3^5 \times 2)$ 、 $L_{32}(8 \times 4^6 \times 2^6)$

每一张同水平的标准正交表都有一张与其相对应的交互作用表(即用于查找任何两列的交互作用所在的列)，例如，与 $L_8(2^7)$ 正交表对应的交互作用表见表4-11。

表 4-11 $L_8(2^7)$ ：二列间的交互作用表

| 列号 | 列号 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|----|------|-----|-----|-----|-----|-----|-----|
| 1 | (1) | 3 | 2 | 5 | 4 | 7 | 6 |
| 2 | | (2) | 1 | 6 | 7 | 4 | 5 |
| 3 | | | (3) | 7 | 6 | 5 | 4 |
| 4 | | | | (4) | 1 | 2 | 3 |
| 5 | | | | | (5) | 3 | 2 |
| 6 | | | | | | (6) | 1 |
| 7 | | | | | | | (7) |

表4-11 的使用方法很简单，例如，希望查找第2列与第5列的交互作用所在的列，行向看“2”，纵向看“5”，交叉处为“7”，即第2列与第5列的交互作用应出现在第7列上。

12. 均匀设计

约定多个试验因素的任何一种水平组合称为一个试验点，则当多个试验因素的某些试验点在空间上分布非常均匀，且达到一定的“标准”时，就称由这样的一些试验点所决定的试验因素的水平组合结构为一个均匀设计。

这里所说的“标准”就是度量一个设计所决定的全部试验点在空间的分布是否均匀的“偏差函数”取值很小或最小。

13. 多因素非平衡组合试验

仅根据自己的专业知识和认识水平，人为确定应做几组试验，以“组别”或“处理”或“试验方案”为全部组的总称。讨论问题或处理资料时，就简单地按“单因素多水平设计”来处置。这种非标准的多因素试验安排被称为“多因素非平衡组合试验(不能称为设计)”。对其资料的处理只能通过拆分处理。

4.2.3 重复测量设计

重复测量设计，即在给予某种处理后，在几个不同的时间点上从同一个受试对象(或样品)上重复获得指标的观测值；有时是从同一个个体的不同部位(或组织)上重复获得指标的观测值。

由于这种设计符合许多生物医学试验本身的特点，故在生物医学科研中应用的频率相当

高。如果试验中共有 k 个试验因素，其中只有 m 个因素与重复测量有关，就称为具有 m 个重复测量的 k 因素设计。

此设计具有一个明显的特点，即在不同条件下，从同一个受试对象身上观测到 k 个数据($k \geq 2$)；测自同一受试者的 k 个数据之间具有不相等的相关性，即间隔越近相关性越强，反之亦然。

1. “时间”为重复测量因素

(1) 具有一个重复测量的单因素设计

试验中只有 1 个试验因素且该试验因素为重复测量因素。

【例 4-13】 某外科医生在每一个病人的手术过程中的 3 个不同阶段测得病人门静脉压力资料见表 4-12，该资料所代表的试验属于具有一个重复测量的单因素设计。

(2) 具有一个重复测量的两因素设计

试验中有 2 个试验因素且其中 1 个试验因素为重复测量因素。

【例 4-14】 资料见表 4-13，每一行中的三个数据测自同一只大鼠，该资料属于具有一个重复测量的两因素设计。

表 4-12 10 名病人在手术过程中的 3 个不同阶段门静脉压力的测定结果

| 病例号 | 门静脉压力 (cmH ₂ O) | | |
|-----|----------------------------|----------------|----------------|
| | A (阶段): | A ₁ | A ₂ |
| 1 | | 40 | 36 |
| 2 | | 19 | 17 |
| ⋮ | | ⋮ | ⋮ |
| 10 | | 36 | 37 |

表 4-13 大鼠分别接受某药三种不同剂量后在三个不同时间点上心率的测定结果

| 药物 剂量 | 鼠号 | 心率 (次/分) | | |
|----------|----|--------------|-----|-----|
| | | 给药后时间 (min): | 0 | 2 |
| 低剂量组 | 1 | | 205 | 246 |
| | 2 | | 313 | 338 |
| | ⋮ | | ⋮ | ⋮ |
| 中剂量组 | 7 | | 277 | 268 |
| | 8 | | 339 | 296 |
| | ⋮ | | ⋮ | ⋮ |
| 高剂量组 | 13 | | 221 | 193 |
| | 14 | | 278 | 245 |
| | ⋮ | | ⋮ | ⋮ |

2. “处理”与“部位”组合起来构成复合型重复测量因素

【例 4-15】 具有两个重复测量的两因素设计实例：用贲门癌患者的标本制成液体，在三种不同处理条件下观测鸡胚背根神经节与鸡胚交感神经节中长出突起的神神经节的比例。现有贲门癌患者 10 例，将每人的标本均分成三份，分别给予三种不同的处理(因素 A)，即 A₁(加入 100ng/mL 神经生长因子)、A₂(加入 200ng/mL 神经生长因子)和 A₃(单用贲门癌培养液)；并对每种处理后的标本中的两种类型神经节(因素 B)，即 B₁(背根神经节)与 B₂(交感神经节)，观测长出突起的神神经节的比例(Y)。设计格式和资料见表 4-14，该资料为具有两个重复测量的两因素设计。

表 4-14 贲门癌患者的标本经三种处理后两种神经节中长出突起的神神经节的比例

| 病例号 | Y (长出突起的神神经节的比例) | | | | | |
|-----|----------------------------------|------------------|----------------------------------|------------------|----------------------------------|------------------|
| | A ₁ (B ₁) | B ₂) | A ₂ (B ₁) | B ₂) | A ₃ (B ₁) | B ₂) |
| 1 | 0.50 | 0.43 | 0.50 | 0.43 | 0.80 | 0.50 |
| 2 | 0.63 | 0.38 | 0.55 | 0.50 | 0.71 | 0.63 |
| 3 | 0.50 | 0.50 | 0.54 | 0.50 | 0.83 | 0.67 |
| 4 | 0.43 | 0.43 | 0.50 | 0.38 | 0.70 | 0.57 |

续表

| 病例号 | Y(长出突起的神经节的比例) | | | | | |
|-----|---------------------------------|------------------|---------------------------------|------------------|---------------------------------|------------------|
| | A ₁ (B ₁ | B ₂) | A ₂ (B ₁ | B ₂) | A ₃ (B ₁ | B ₂) |
| 5 | 0.50 | 0.40 | 0.50 | 0.57 | 0.70 | 0.71 |
| 6 | 0.44 | 0.38 | 0.63 | 0.50 | 0.77 | 0.63 |
| 7 | 0.43 | 0.50 | 0.50 | 0.38 | 0.69 | 0.50 |
| 8 | 0.63 | 0.44 | 0.40 | 0.38 | 0.70 | 0.50 |
| 9 | 0.44 | 0.50 | 0.50 | 0.44 | 0.67 | 0.50 |
| 10 | 0.44 | 0.43 | 0.45 | 0.40 | 0.60 | 0.75 |

3.“处理”与“时间”组合起来构成复合型重复测量因素

【例 4-16】 具有两个重复测量的三因素设计实例：某研究者为了研究阿托品累计用量对有机磷药物中毒后的治疗效果，将 10 只大鼠随机均分成两组，第一组 5 只接受有机磷农药使之中毒，第二组 5 只未接受有机磷农药。两组大鼠均给予一次乙酰胆碱，同时，4 次给予阿托品治疗，每次均为 2mg/kg，在每次给予阿托品治疗后的 2min、5min、10min 观测膈肌对电刺激产生的反应，用“面积”表示，面积越大，治疗效果越好。设计和资料见表 4-15，该资料为具有两个重复测量的三因素设计。

表 4-15 连续 4 次给阿托品治疗不同时间点上膈肌对电刺激产生反应的观测结果

| 中毒与否(A) | 大鼠编号 | 面 积 | | | | | | |
|---------|------|--------|-------------------|------|------|-------------------|------|------|
| | | B 与 C: | B ₁ (2 | 5 | 10) | B ₂ (2 | 5 | 10) |
| 是 | 1 | | 48.0 | 48.6 | 51.0 | 52.8 | 50.6 | 49.7 |
| | 2 | | 45.1 | 47.7 | 48.7 | 47.1 | 45.7 | 42.3 |
| | 3 | | 45.1 | 47.7 | 48.7 | 46.0 | 46.0 | 43.7 |
| | 4 | | 41.4 | 41.1 | 42.0 | 39.1 | 38.0 | 35.4 |
| | 5 | | 57.7 | 60.6 | 61.4 | 58.6 | 56.8 | 52.0 |
| 否 | 1 | | 64.8 | 66.3 | 67.1 | 66.0 | 65.1 | 62.0 |
| | 2 | | 59.7 | 59.7 | 59.1 | 59.4 | 59.1 | 59.4 |
| | 3 | | 53.1 | 54.6 | 54.6 | 52.3 | 51.7 | 49.1 |
| | 4 | | 61.1 | 62.9 | 60.6 | 56.8 | 52.8 | 47.1 |
| | 5 | | 84.6 | 85.1 | 84.3 | 80.3 | 77.1 | 72.3 |

注：表中右边还有 6 列数据未列出，其表头为“B₃ (2 5 10) min”和“B₄ (2 5 10) min”。

4.3 比较类型概述

在临床试验中，按照研究目的可分为差异性试验、优效性试验、等效性试验和非劣效性试验，与之相对应的对试验结果的比较类型称为差异性检验、优效性检验、等效性检验和非劣效性检验。

4.3.1 四种比较类型的概念

1. 差异性检验

差异性检验是指主要研究目的为显示两种治疗效果之间的差别大小与 0 之间的差别是否具有统计学意义的试验。在试验设计阶段不需要设定任何界值。

2. 等效性检验

等效性检验是指主要研究目的是显示两种治疗效果间差别的大小在临床上并无重要性的试验。在试验设计阶段需要设定等效性界值(δ_L, δ_U)来界定两种药物的等效性。

3. 优效性检验

优效性检验是指主要研究目的是显示所研究药物的治疗效果优于对照药(阳性或安慰剂对照)的试验。在试验设计阶段需要设定一个界值 δ_U ,来界定试验药的优效性。

4. 非劣效性检验

非劣效性检验是指主要研究目的是显示试验药的治疗效果在临床上不差于(非劣于)阳性对照药的试验。在试验设计阶段需要设定一个界值 δ_L ,来界定试验药是否不比阳性对照药疗效差过预先设定的这个界值。

4.3.2 四种比较类型下检验假设及结论的正确陈述

1. 三种差异性检验的检验假设及结论的正确陈述

(1) 双侧差异性检验

检验假设:

$H_0: \mu_T - \mu_R = 0$ 试验组与对照组的总体平均数相等。

$H_1: \mu_T - \mu_R \neq 0$ 试验组与对照组的总体平均数不等。

$\alpha = 0.05$

结论陈述:

当 $p > \alpha$ 时, 不能拒绝 H_0 , 还不能认为试验组与对照组平均数之间的差别有统计学意义, 即还不能认为试验组与对照组的平均数不等。

当 $p \leq \alpha$ 时, 拒绝 H_0 , 接受 H_1 , 认为试验组与对照组的平均数之间的差别有统计学意义, 即认为试验组与对照组的平均数不等。

(2) 左单侧差异性检验

检验假设:

$H_0: \mu_T - \mu_R \geq 0$ 试验组的总体平均数大于或等于对照组的总体平均数。

$H_1: \mu_T - \mu_R < 0$ 试验组的总体平均数小于对照组的总体平均数。

$\alpha = 0.05$

结论陈述:

当 $p > \alpha$ 时, 不能拒绝 H_0 , 还不能认为试验组的平均数小于对照组的平均数。

当 $p \leq \alpha$ 时, 拒绝 H_0 , 接受 H_1 , 认为试验组的平均数小于对照组的平均数。

(3) 右单侧差异性检验

检验假设:

$H_0: \mu_T - \mu_R \leq 0$ 试验组的总体平均数小于或等于对照组的总体平均数。

$H_1: \mu_T - \mu_R > 0$ 试验组的总体平均数大于对照组的总体平均数。

$\alpha = 0.05$

结论陈述:

当 $p > \alpha$ 时, 不能拒绝 H_0 , 还不能认为试验组的平均数大于对照组的平均数。

当 $p \leq \alpha$ 时, 拒绝 H_0 , 接受 H_1 , 认为试验组的平均数大于对照组的平均数。

2. 非劣效性检验的检验假设及结论的正确陈述

检验假设:

$H_0: \mu_T - \mu_R \leq \delta_L$ 试验组的总体平均数与对照组的总体平均数之差小于或等于 δ_L , 即试验组劣效于对照组。

$H_1: \mu_T - \mu_R > \delta_L$ 试验组的总体平均数与对照组的总体平均数之差大于 δ_L , 即试验组非劣效于对照组。

$\alpha = 0.05$

结论陈述:

当 $p > \alpha$ 时, 不能拒绝 H_0 , 说明试验组的总体平均数与对照组的总体平均数之差小于或等于 δ_L , 还不能认为试验组非劣效于对照组。

当 $p \leq \alpha$ 时, 拒绝 H_0 , 接受 H_1 , 说明试验组的总体平均数与对照组的总体平均数之差大于 δ_L , 认为试验组非劣效于对照组。

3. 等效性检验的检验假设及结论的正确陈述

检验假设:

$H_{0(1)}: \mu_T - \mu_R \leq \delta_L$ 试验组的总体平均数与对照组的总体平均数之差小于或者等于 δ_L 。

$H_{1(1)}: \mu_T - \mu_R > \delta_L$ 试验组的总体平均数与对照组的总体平均数之差大于 δ_L 。

$\alpha' = 0.025$ (单侧)

$H_{0(2)}: \mu_T - \mu_R \geq \delta_U$ 试验组的总体平均数与对照组的总体平均数之差大于或者等于 δ_U 。

$H_{1(2)}: \mu_T - \mu_R < \delta_U$ 试验组的总体平均数与对照组的总体平均数之差小于 δ_U 。

$\alpha' = 0.025$ (单侧)

结论陈述:

当 $p_1 > \alpha'$ 或者 $p_2 > \alpha'$ 时, 还不能认为试验组与对照组等效。

当 $p_1 \leq \alpha'$ 且 $p_2 \leq \alpha'$ 时, 可以认为试验组与对照组等效。

4. 优效性检验的检验假设及结论的正确陈述

检验假设:

$H_0: \mu_T - \mu_R \leq \delta_U$ 试验组的总体平均数与对照组的总体平均数之差小于或者等于 δ_U 。

$H_1: \mu_T - \mu_R > \delta_U$ 试验组的总体平均数与对照组的总体平均数之差大于 δ_U 。

$\alpha = 0.05$

结论陈述:

当 $p > \alpha$ 时, 不能拒绝 H_0 , 说明试验组的总体平均数与对照组的总体平均数之差小于或等于 δ_U , 还不能认为试验组优效于对照组。

当 $p \leq \alpha$ 时, 拒绝 H_0 , 接受 H_1 , 说明试验组的总体平均数与对照组的总体平均数之差大于 δ_U , 认为试验组优效于对照组。

4.3.3 合理选择临床试验的比较类型

1. 选择差异性检验的理由与场合

(1) 选择双侧差异性检验的理由与场合

当预试验的结果表明, 试验药与对照药效果接近, 而且两者之间的效应指标的差量取正值

还是负值尚不能确定时,研究目的是考察试验药与对照药疗效之差量与0之间的差别是否具有统计学意义(没有设定有临床意义的界值),选用双侧差异性检验。

(2) 选择左单侧差异性检验的理由与场合

当预试验的结果表明,试验药的疗效总比对照药的疗效稍差一些,但差量并非足够大时,研究目的是考察试验药与对照药疗效之差量是否具有统计学意义(没有设定有临床意义的界值),选用左单侧差异性检验。

(3) 选择右单侧差异性检验的理由与场合

当预试验的结果表明,试验药的疗效总比对照药的疗效稍好一些,但差量并非足够大时,研究目的是考察试验药与对照药疗效之差量是否具有统计学意义(没有设定有临床意义的界值),选用右单侧差异性检验。

2. 选择非劣效性检验的理由与场合

当预试验的结果表明,虽然试验药的效果比对照药效果略差一些,但两者之间的效应指标的差量在数量上并非足够大时,结合临床专业知识可知,此差量尚达不到具有临床实际意义的界值。此时,为显示试验药的治疗效果在临床上不比阳性对照药差,选用非劣效性检验。在试验设计阶段设定了一个有临床意义的界值 δ_L ,来界定试验药是否不比阳性对照药疗效差过预先设定的这个界值。

3. 选择等效性检验的理由与场合

当预试验的结果表明,试验药的效果不会比对照药效果好很多,也不会差很多,而且两者之间的效应指标的差量的绝对值在数量上并非足够大时,结合临床专业知识可知,此差量的绝对值将小于具有临床实际意义的界值。此时,为显示两种治疗药物之间的差别大小在临床上并无重要意义,选用等效性检验。在试验设计阶段设定了两个有临床意义的等效性界值(δ_L , δ_U)来界定两种治疗的等效性。

4. 选择优效性检验的理由与场合

当预试验的结果表明,试验药的效果不仅比对照药效果好,而且,两者之间的效应指标的差量在数量上相当可观时,结合临床专业知识可知,此差量具有临床上的实际意义。此时,为了通过正式临床试验,显示试验药的治疗效果优效于对照药(安慰剂对照或阳性对照),选用优效性检验。在试验设计阶段设定了一个有临床意义的界值 δ_U ,来界定试验药的优效性。

4.4 样本含量与检验效能估计概述

目前,绝大多数研究工作者不是通过科学的样本含量估计确定样本量,而是自己随意设定样本量。当假设检验的结果为“阴性”,即不能拒绝原假设时,也很少有研究者去分析假设检验的检验效能,进而判断“阴性”结果的可信度如何。造成这种状况的原因:一是研究者没有认识到样本含量估计与检验效能分析对科学研究的重要性,二是研究者不知道如何进行样本含量估计与检验效能分析。与样本含量与检验效能估计有关的具体方法和 SAS 实现将在本书第6章具体讲述。本节简要介绍样本含量与检验效能估计的概念、意义与作用。

4.4.1 样本含量估计的概念、意义与作用

样本含量,即观察例数的多少,在数理统计上习惯于称为样本大小。在一个科研项目中,所

需样本含量的大小与许多因素有关,涉及统计研究设计的类型(调查设计、试验设计和临床试验设计),对结果精确度要求的高低和研究进度的快慢以及资料的性质和研究目的等,需借助适当的公式,进行样本含量的估计。研究者可以根据需要和可能来确定一个适合的样本含量。

根据假设检验的原理,若样本太小,会使本来存在的差别不能显示出来,难以获得正确的研究结果,结论也缺乏充分的依据;反之,若样本太大,也会增加实际工作中的困难,不必要地浪费人力、物力、财力和时间。此外,由于样本太大,可能投入不足,科研过程的质量控制下降,从而引入不必要的干扰因素,对研究结果造成不良影响。

4.4.2 检验效能估计的概念、意义与作用

检验效能,又称为把握度 $\text{power} = 1 - \beta$ 。其意义是,当所研究的总体与原假设确有差别时,按检验水平 α 能够发现它(拒绝原假设)的概率。如果 $1 - \beta = 0.09$,则意味着当原假设不成立时,理论上在每 100 次抽样中,在 α 的检验水准上平均有 90 次能够拒绝原假设。一般情况下对同一检验水准 α ,检验效能大的检验方法更可取。一般情况下要求检验功效应在 0.75 以上,否则检验的结果很可能反映不出总体的真实差别,出现非真实的阴性结果。

4.5 随机化方法概述

随机原则是统计设计中应该遵循的基本原则之一。它是指采用随机的方式来选取和分配受试对象或观察对象或调查对象,即研究总体中每个个体都具有同等的机会被抽中进入样本,而样本中每个受试对象都有同等的机会被分配到各个试验组中。本节将简要介绍随机化概念、意义与作用,与随机化有关的具体内容和 SAS 实现将在第 7 章详细介绍。

4.5.1 随机化的概念

随机化主要包括两个方面。

(1)随机抽样:每一个符合条件的试验对象参加试验的机会相同,即总体中每个个体有相同的会被抽取进入样本之中。

(2)随机分组:每个试验对象被分配到不同组(通常为对照组、不同处理组)中去的机会相同。

4.5.2 随机化的意义与作用

随机抽样可以保证所得到的样本具有代表性,使试验结论具有普遍意义;随机分组可以保证各组间试验对象尽可能均衡一致,以提高各组间的可比性。

4.5.3 随机抽样方法

常用的随机抽样方法有:单纯随机抽样、系统抽样、整群抽样、分层抽样以及多阶段抽样。

(1)单纯随机抽样:又称简单随机抽样,它是利用随机的方法从总体中抽取部分抽样单位作为样本,总体中每个抽样单位被抽中的概率相等。单纯随机抽样可以采用掷硬币、掷骰子、抽签或者查随机数字表的方法,也可以在计算机上利用 SAS、SPSS 等软件来实现。

(2)系统抽样:也称等距抽样或机械抽样,它是按照一定的顺序对总体中各抽样单位进行排序,根据样本大小确定抽样间隔,然后随机确定起点,按照抽样间隔抽取每一个单位。它也是在随机抽样的基础上进行的,即要求每次抽样的起点必须是随机的。

(3)整群抽样:又称集团抽样,它是先将总体分成 k 个群(组),从中随机抽取 k 个群,然

后把抽取的群中的所有观察单位组成样本。

(4) 分层抽样: 分层抽样是先按对观测指标影响较大的某种特征, 将总体分为若干个层, 再从每一层内随机抽取一定数量的抽样单位组成样本。分层原则是使层内差异尽可能小, 层间的差异尽可能大, 从而提高调查的精度。分层抽样按照各层之间的抽样比是否相同, 可分为等比例抽样与非等比例抽样。

(5) 多阶段抽样: 又称多级抽样, 它是指抽样过程分为多个阶段来进行, 每个阶段可以采取不同的抽样方法。多阶段抽样先从总体中抽取一级抽样单元, 再从中抽取范围较小的二级抽样单元, 然后依次抽取更次级的抽样单元, 直到获得所需要的样本量时抽样才结束。

4.5.4 随机分组方法

随机分组方法包括完全随机、分层随机、区组随机、分层区组随机、动态随机。

(1) 完全随机化是指直接对受试对象进行随机化分组, 在事先或者实施过程中不做任何限制和干预或调整。完全随机化分组后, 各组受试对象的例数不一定相等。

(2) 分层随机化是指在随机化过程中, 首先依据可能影响试验过程和结果的重要非试验因素(如年龄、性别、病情、疾病分期等)对受试对象进行分层, 然后在每一层内进行完全随机化分组, 即把每一层中的受试对象完全随机地均分入试验组和对照组中去。分层的目的是使某些对结果影响较大的非试验因素在各组(对照组与试验组)中的影响尽可能相等。

(3) 区组随机化也称均衡随机化或限制性随机化, 它是根据研究对象进入试验的时间顺序, 将全部病例分为例数相同的若干区组, 每一区组内病例随机分配到各研究组。这种随机化方法比较适合临床科研中入选患者分散就诊的特点, 可以确保整个试验期间进入每一组的对象数基本相等, 同一区组中各组病人数相等, 区组大小(长度)是处理数的倍数, 长度一般取 6~10 为好。

(4) 分层区组随机化是按照事先确定的分层因素即重要的非试验因素划分形成层, 在层内设置区组, 随后根据受试者的分层因素将其分入区组, 在区组内随机分配, 保证每个试验组的受试者数相同且入组受试者的分层因素特征尽可能相近。在多中心临床试验中, 目前普遍推荐使用该方法, 以中心为分层因素。

(5) 动态随机化是“按不平衡指数最小的分配原则”分组, 适合用于临床试验研究中。因为患者来医院就诊, 在患病严重程度、患病时间等重要非试验因素方面不一定是均衡的, 可能带有某种程度上的“聚集性”, 为避免此“聚集性”造成试验组和对照组间不均衡, 可采用动态随机化。

4.6 本章小节

统计设计包括试验设计、临床试验设计和调查设计三类。统计设计的要点是三要素、四原则、设计类型和质量控制。

设计类型包括单组设计、配对设计、成组设计、单因素 k 水平设计; 随机区组设计、析因设计、嵌套设计、裂区设计、拉丁方设计、交叉设计、正交设计、均匀设计; 重复测量设计等。

临床试验中的比较类型包括差异性检验、等效性检验、优效性检验与非劣效性检验。

样本含量与检验效能估计与科研结论可信性密切相关。

随机化包括随机抽样和随机分组。常用的随机抽样方法有: 单纯随机抽样、系统抽样、整群抽样、分层抽样以及多阶段抽样; 随机分组方法包括完全随机、分层随机、区组随机、分层区组随机、动态随机。

(胡良平 高辉 陶丽新 胡完)

第 5 章 构建设计类型的 SAS 实现

医学统计学中，试验设计类型较多，不同设计类型资料的呈现方式不同，正确的呈现方式有利于正确了解实验所用设计类型，并且，有利于正确记录实验结果，从而有利于正确选用统计分析方法。本章主要讲述常用多因素设计类型的标准列表格式以及 SAS 实现。

5.1 常用标准多因素设计类型的列表格式

当实验中涉及两个或两个以上因素时，如何以统计表的形式将原始数据全部表达出来呢？本节将通过具体的实例，以统计表的形式呈现多种多因素设计下收集的原始定量资料。

5.1.1 随机区组设计

【例 5-1】 某研究所研制了三个降血脂中药复方制剂，现拟对三个复方制剂与标准降脂药（安妥明）的疗效进行比较。选取品种相同、健康的雄性家兔 16 只，按其体重大小分为四个随机区组（每个区组内的 4 只动物体重最接近），各药物组的动物均饲以同样高脂饮食，并每日分别灌以不同药物，第 45 天处死动物，观察其冠状动脉根部动脉粥样硬化斑块大小，资料见表 5-1。

表 5-1 用四种降脂药物时动物的冠状动脉粥样硬化斑块面积

| 体重 分组 | 斑块面积 (cm ²) | | | | |
|----------|-------------------------|-------|-------|-------|-------|
| | 药物： | 安妥明组 | 降脂甲方组 | 降脂乙方组 | 降脂丙方组 |
| 1 | | 0.000 | 0.283 | 0.114 | 0.094 |
| 2 | | 0.009 | 0.196 | 0.146 | 0.131 |
| 3 | | 0.003 | 0.217 | 0.158 | 0.065 |
| 4 | | 0.001 | 0.236 | 0.159 | 0.087 |

5.1.2 含一个协变量的随机区组设计

【例 5-2】 为了研究 5 种品种（A ~ E）玉米的产量之间的差别是否有统计学意义，农业科技人员做了如下的实验：将一块实验田划分成 4 个区组，再将每个区组等分成 5 个子区，用随机的方法决定每个区组内 5 个子区中任何一个种植一种玉米。该实验由于气候不良，有许多子区内有缺株，各区株数参差不齐。好在研究者记录下了每个子区的存活株数（即协变量）及产量，资料见表 5-2。

表 5-2 5 种不同玉米在各子区的存活株数及产量

| 区 组 | 株数 产量 | | 株数 产量 | | 株数 产量 | | 株数 产量 | | 株数 产量 | |
|--------|-------|----|-------|----|-------|----|-------|----|-------|----|
| | A | | B | | C | | D | | E | |
| 1 | 10 | 18 | 12 | 36 | 17 | 40 | 14 | 21 | 12 | 42 |
| 2 | 8 | 17 | 13 | 28 | 15 | 36 | 14 | 23 | 10 | 36 |
| 3 | 6 | 14 | 8 | 27 | 13 | 35 | 17 | 24 | 10 | 38 |
| 4 | 8 | 15 | 11 | 30 | 11 | 29 | 15 | 20 | 16 | 52 |

5.1.3 平衡不完全随机区组设计

【例 5-3】 用四种方法治疗脚气，受试者为两脚都患脚气的患者，每只脚接受一种处理，观察指标为治疗效果评分。设计格式和资料见表 5-3。

表 5-3 四种方法治疗脚气的评分

| 受试者 编号 | 疗 效 评 分 | | | | |
|-----------|---------|---|---|---|---|
| | 治疗方法: | 甲 | 乙 | 丙 | 丁 |
| 1 | | 5 | 3 | — | — |
| 2 | | — | — | 4 | 7 |
| 3 | | 2 | — | 6 | — |
| 4 | | — | 4 | — | 8 |
| 5 | | 3 | — | — | 7 |
| 6 | | — | 2 | 5 | — |

5.1.4 拉丁方设计

【例 5-4】 观察 A、B、C 三种中药的促凝作用，以生理盐水作对照(D)，假定每种药物作用的时间很短，观测指标的值很快便能恢复到原先的水平，即可以用同一只动物重复做此实验。现取品种相同、体重接近的雄性白兔 4 只，每只白兔以不同的顺序接受不同的药物，以血浆复钙凝固时间(s，即秒)为指标，设计和资料见表 5-4，此为 4×4 拉丁方设计。

表 5-4 不同中药组测得的血凝时间

| 白兔 编号 | 中药代号(血凝时间(s)) | | | |
|----------|---------------|--------|--------|--------|
| | 实验顺序号: 1 | 2 | 3 | 4 |
| 1 | C(85) | D(102) | A(90) | B(100) |
| 2 | D(108) | A(88) | B(110) | C(80) |
| 3 | A(95) | B(98) | C(82) | D(110) |
| 4 | B(105) | C(78) | D(104) | A(92) |

5.1.5 交叉设计

【例 5-5】 为研究 A(90402 中药复方)、B(安慰剂)两种处理对提高高原劳动能力的影
响，以条件近似的 10 名健康人作为受试对象，把受试对象和测定顺序(冬季、春季)作为两个重要的非处理因素，每人都服用两种药物各一次。实验设计方案如下：将条件最接近的每两人配成一对，用随机的方法确定每对中之一使用 A、B 药物的顺序，另一个人服药的顺序相反。设计格式和实验结果列在表 5-5 中，其观测指标为 PWC170(即把心跳校正到 170 次/min 时能做的功)，此为二阶段配对交叉设计。

表 5-5 两种药物对提高高原劳动能力的实验结果

| 受试者 编号 | 药物与 PWC170 | | | 受试者 编号 | 药物与 PWC170 | | |
|-----------|------------|-------|---------|-----------|------------|-------|---------|
| | 实验时间: | 冬季 | 春季 | | 实验时间: | 冬季 | 春季 |
| 1 | A | 159.4 | B 153.8 | 6 | B | 129.9 | A 146.1 |
| 2 | B | 129.4 | A 159.8 | 7 | B | 166.3 | A 210.9 |
| 3 | B | 122.1 | A 137.6 | 8 | A | 208.7 | B 169.9 |
| 4 | A | 130.4 | B 137.8 | 9 | A | 180.1 | B 150.7 |
| 5 | A | 150.7 | B 140.2 | 10 | B | 143.4 | A 150.8 |

注：(1, 2), ..., (9, 10)号分别为配对组；A(90402 中药复方)、B(安慰剂)。

5.1.6 无重复实验的双因素设计

【例 5-6】 某研究者研究两种来源的家兔在 4 种室温条件下的血糖值，其定量资料可以表达成表 5-6 的形式。其中，家兔来源和室温都是实验因素，仅当两因素之间的交互作用可以忽略不计且特定条件下的实验结果相对比较稳定时，两因素各水平组合下才可以不做重复实验。否则，必须做重复实验(此时，通常就转化为两因素析因设计了)。

表 5-6 不同来源和不同室温对家兔血糖值的影响情况

| 家兔来源 | 血糖值(毫克%) | | | | | |
|------|----------|------|------|------|-----|------|
| | 室温(℃): | 5 | 10 | 15 | 20 | 25 |
| I | | 14.0 | 12.0 | 11.0 | 8.2 | 8.2 |
| II | | 16.0 | 14.0 | 10.0 | 8.3 | 11.0 |

5.1.7 嵌套设计

【例 5-7】 在某城市调查某种疾病的患病率，调查时在地区因素 A 上分为 3 水平：A₁(工厂区)、A₂(文化区)、A₃(市区)；在年龄因素 B 上分为 2 水平：B₁(儿童)、B₂(成人)；性别因素 C 上有 2 个水平：C₁(男性)、C₂(女性)。三因素各水平搭配下都重复两次调查，每次都调查了足够数量的人。假定由专业知识得知：对该病的影响，地区的作用最大、年龄次之、性别最小，数据见表 5-7，该资料所对应的实验设计类型为系统分组(或嵌套)设计。

表 5-7 地区、年龄、性别对某种病患病率的影响

| 因素 B | 因素 C | 患病率(%) | | | | | | |
|----------------|----------------|--------|----------------|----------------|----------------|-----|------|------|
| | | 因素 A: | A ₁ | A ₂ | A ₃ | | | |
| B ₁ | C ₁ | | 12.4 | 15.6 | 10.5 | 6.4 | 17.3 | 14.4 |
| | C ₂ | | 13.1 | 10.2 | 9.4 | 7.6 | 16.9 | 15.8 |
| B ₂ | C ₁ | | 8.2 | 15.1 | 1.3 | 4.5 | 13.0 | 11.7 |
| | C ₂ | | 7.4 | 10.7 | 7.2 | 5.1 | 14.5 | 12.3 |

5.1.8 裂区设计

【例 5-8】 为研究不同瘤株的生瘤效果和不同浓度蛇毒的抑瘤作用，先将 48 只小鼠按原始体重分成 3 个随机区组(注：每个区组内的 16 只小鼠体重最接近)，将每个区组内的 16 只小鼠随机地均分成 4 组，分别接种 4 种不同的瘤株，观察肿瘤生长情况。1d 后再对接同一种瘤株的 4 只小鼠分别腹腔注射 4 种不同浓度的蛇毒，连续用蛇毒抑瘤 10d，停药 1d 后解剖测瘤重。设计格式和资料见表 5-8，此为裂区(或分割)设计。

表 5-8 不同瘤株与不同浓度的蛇毒共同作用后对小鼠抑瘤效果的影响

| 瘤株 A | 蛇毒浓度 B | 瘤重(g) | | | |
|-----------------------|----------------|-------|------|------|------|
| | | 随机区组: | 1 | 2 | 3 |
| A ₁ (S180) | B ₁ | | 0.80 | 0.76 | 0.36 |
| | B ₂ | | 0.36 | 0.26 | 0.31 |
| | B ₃ | | 0.17 | 0.28 | 0.16 |
| | B ₄ | | 0.28 | 0.13 | 0.11 |

续表

| 瘤株 A | 蛇毒浓度 B | 瘤重 (g) | | | |
|----------------------|----------------|--------|------|------|------|
| | | 随机区组: | 1 | 2 | 3 |
| A ₂ (HS) | B ₁ | | 0.74 | 0.43 | 0.57 |
| | B ₂ | | 0.50 | 0.46 | 0.32 |
| | B ₃ | | 0.42 | 0.20 | 0.20 |
| | B ₄ | | 0.36 | 0.26 | 0.32 |
| A ₃ (EC) | B ₁ | | 0.31 | 0.55 | 0.32 |
| | B ₂ | | 0.20 | 0.15 | 0.20 |
| | B ₃ | | 0.38 | 0.18 | 0.26 |
| | B ₄ | | 0.25 | 0.21 | 0.14 |
| A ₄ (ARS) | B ₁ | | 0.48 | 0.57 | 0.33 |
| | B ₂ | | 0.18 | 0.30 | 0.29 |
| | B ₃ | | 0.44 | 0.27 | 0.27 |
| | B ₄ | | 0.22 | 0.30 | 0.37 |

5.1.9 析因设计

【例 5-9】 某医院用中药复方治疗高胆固醇血症，把 12 例高胆固醇患者随机分为四组，用不同疗法治疗。第一组用一般疗法，第二组在一般疗法上外加用甲药，第三组在一般疗法上外加用乙药，第四组在一般疗法上外加用甲药和乙药，一个月后观察胆固醇降低数 (mg%) 资料如表 5-9 所示，此为两因素 (或 2×2) 析因设计。

表 5-9 甲、乙两药治疗高胆固醇血症的疗效

| 甲药使用与否 | 胆固醇降低值 (mg%) | | | | | | |
|--------|--------------|----|----|----|----|----|----|
| | 乙药使用与否: | 不用 | | | 用 | | |
| 不用 | | 16 | 25 | 18 | 28 | 31 | 23 |
| 用 | | 56 | 44 | 42 | 64 | 78 | 80 |

5.1.10 含区组因素的析因设计

【例 5-10】 目的：研究小鼠在以不同注射剂量 (每天总注射量，因素 A) 和不同注射频率 (每天注射次数，因素 B) 下药剂 ACTH 对尿总酸度的影响。因素 A 有 3 个水平：A₁、A₂、A₃ 是三种不同的日注射量；B 有两个水平：B₁ (每天注射次数很少)、B₂ (每天注射次数很多)。这两个因素共有 6 种水平组合，称为 6 个处理组。将 24 只小鼠按某些条件 (即重要的非处理因素) 分为 4 个随机区组，每个随机区组中的 6 只小鼠在规定的条件方面最为接近。然后，将每个随机区组中的 6 只小鼠随机地分入 6 个处理组中去。设计格式和实验数据见表 5-10，该资料所对应的实验设计类型为含区组因素的析因设计。

表 5-10 A、B 两因素对尿总酸度影响的实验结果

| 随机区组编号 | 尿总酸度 (单位) | | | | | |
|--------|----------------------------------|------------------|----------------------------------|------------------|----------------------------------|------------------|
| | A ₁ (B ₁) | B ₂) | A ₂ (B ₁) | B ₂) | A ₃ (B ₁) | B ₂) |
| 1 | 33.6 | 33.0 | 33.0 | 28.5 | 31.4 | 30.7 |
| 2 | 37.1 | 30.5 | 29.5 | 31.8 | 28.3 | 28.2 |
| 3 | 34.1 | 33.3 | 29.2 | 29.9 | 28.9 | 28.4 |
| 4 | 34.6 | 34.4 | 30.7 | 28.3 | 28.6 | 30.6 |

5.1.11 正交设计

【例 5-11】 在乙酰苯胺磺化工工艺的研究中，有 4 个因素：反应温度(℃)A、反应时间(h)B、硫酸浓度(%)C、操作方法 D，各取 2 水平，实验结果为产物的“收率(%)”。实验目的：希望弄清各因素在怎样的搭配条件下“收率”最高。已知反应温度与反应时间之间的交互作用不可忽视，各实验条件下不必进行重复实验，希望总实验次数尽可能少一些。4 个实验因素的水平分别如下：

| | | |
|----------|------|------|
| 因素名称(单位) | 1 水平 | 2 水平 |
| 反应温度(℃) | 50 | 70 |
| 反应时间(h) | 1 | 2 |
| 磺酸浓度(%) | 17 | 27 |
| 操作法 | 搅拌 | 不搅拌 |

选择 $L_8(2^7)$ 正交表是非常合适的，因为该正交表的自由度为 7，比所需要的自由度 5 多 2 个自由度，即正交表中将有两个空列，可用于估计实验误差，此时，只需要做 8 次实验，就可以满足研究者的要求。

设因素 A 为反应温度(℃)、因素 B 为反应时间(h)、因素 C 为磺酸浓度(%)、因素 D 为操作法。选用 $L_8(2^7)$ 正交表，将 A 放在第 1 列、B 放在第 2 列、查找 $L_8(2^7)$ 正交表的交互作用表，得交互作用 $A \times B$ 应落在第 3 列，于是，可把 C 放在第 4 列，D 可以放在 5、6、7 三列中任何一列上，不妨将 D 放在第 7 列上，见表 5-11。

表 5-11 $L_8(2^7)$ 正交表

| 实验号 | 列号: 1 (A) | 2 (B) | 3 (A×B) | 4 (C) | 5 | 6 | 7 (D) | Y(实验结果) |
|-----|-----------|-------|---------|-------|---|---|-------|---------|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | X |
| 2 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | X |
| 3 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | X |
| 4 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | X |
| 5 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | X |
| 6 | 2 | 1 | 2 | 2 | 1 | 2 | 1 | X |
| 7 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | X |
| 8 | 2 | 2 | 1 | 2 | 1 | 1 | 2 | X |

根据表 5-11 并不便于进行实验，最好将表中第 1、2、4、7 列抽出来，以便使表格简化，不易看串行。不仅如此，还应将表中的“代码水平”转化成“真实水平”，这样，做实验时才不会出错。结果见表 5-12。实验结束后将实验结果填入表中最后一列。

表 5-12 用 $L_8(2^7)$ 正交表安排 4 个实验因素

| 实验号 | 反应温度(℃) | 反应时间(h) | 磺酸浓度(%) | 操作法 | Y(收率, %) |
|-----|---------|---------|---------|-----|----------|
| 1 | 50 | 1 | 17 | 搅拌 | 65 |
| 2 | 50 | 1 | 27 | 不搅拌 | 74 |
| 3 | 50 | 2 | 17 | 不搅拌 | 71 |
| 4 | 50 | 2 | 27 | 搅拌 | 73 |
| 5 | 70 | 1 | 17 | 不搅拌 | 70 |
| 6 | 70 | 1 | 27 | 搅拌 | 73 |
| 7 | 70 | 2 | 17 | 搅拌 | 62 |
| 8 | 70 | 2 | 27 | 不搅拌 | 67 |

5.1.12 均匀设计

【例 5-12】 某化工厂为提高某化工产品的转化率，经分析研究，决定选取的因素与水平如下：

| 水平 | 反应温度(℃)A | 反应时间(min)B | 用碱量(%)C |
|----|----------|------------|---------|
| 1 | 80 | 90 | 5 |
| 2 | 85 | 120 | 6 |
| 3 | 90 | 150 | 7 |

现欲通过实验找出影响指标转化率(越高越好)的各因素最佳水平组合及各因素的主次顺序，通过预实验发现，因素之间的交互作用可以忽略不计，希望实验次数尽可能少一些。

这里用 $U_9(3^3)$ 均匀表进行设计，见表 5-13。

若将各列上因素的“代码水平”转换为“真实水平”，使设计看上去简单明了，不易出错，见表 5-14。等实验结束后将实验结果填入表中最后一列。

表 5-13 用 $U_9(3^3)$ 均匀表安排三个实验因素

| 实验号 | 例号: 1(A) | 2(B) | 3(C) | Y(实验结果) |
|-----|----------|------|------|---------|
| 1 | 1 | 1 | 2 | X |
| 2 | 1 | 2 | 1 | X |
| 3 | 1 | 3 | 3 | X |
| 4 | 2 | 1 | 3 | X |
| 5 | 2 | 2 | 2 | X |
| 6 | 2 | 3 | 1 | X |
| 7 | 3 | 1 | 1 | X |
| 8 | 3 | 2 | 3 | X |
| 9 | 3 | 3 | 2 | X |

表 5-14 用 $U_9(3^3)$ 均匀表安排三个实验因素

| 实验号 | 反应温度 (℃) | 反应时间 (min) | 用碱量 (%) | Y(转化率, %) |
|-----|-------------|---------------|------------|--------------|
| 1 | 80 | 90 | 6 | 31 |
| 2 | 80 | 120 | 5 | 54 |
| 3 | 80 | 150 | 7 | 38 |
| 4 | 85 | 90 | 7 | 53 |
| 5 | 85 | 120 | 6 | 49 |
| 6 | 85 | 150 | 5 | 42 |
| 7 | 90 | 90 | 5 | 57 |
| 8 | 90 | 120 | 7 | 62 |
| 9 | 90 | 150 | 6 | 64 |

当然，若实验做起来不是特别困难或费用不是特别昂贵，最好能多做几次实验，可以选用 $U_{12}(3^3)$ 均匀表或 $U_{15}(3^3)$ 均匀表。

5.1.13 重复测量设计

【例 5-13】 资料见表 5-15，本实验涉及“性别”和“时间”两个实验因素，其中“时间”是一个重复测量因素，本实验的设计类型为具有一个重复测量的两因素设计。

表 5-15 每只狗服药后在 3 个不同时间点上血中药浓度的测定结果

| 性 别 | 狗 号 | 血药浓度(r/mL) | | | |
|-----|-----|------------|-------|-------|------|
| | | 时间(h) : | 6 | 8 | 24 |
| 公 | 1 | | 169.8 | 137.2 | 31.9 |
| | 2 | | 158.4 | 133.2 | 18.7 |
| | 3 | | 142.8 | 126.6 | 18.1 |
| 母 | 4 | | 131.8 | 130.3 | 17.5 |
| | 5 | | 118.0 | 124.5 | 18.7 |
| | 6 | | 120.8 | 123.5 | 24.8 |

5.2 常用标准多因素设计类型的 SAS 输出格式

5.2.1 如何用 SAS 实现随机区组设计

【例 5-14】 某研究者欲研究注射不同剂量的雌激素对大鼠子宫重量的影响，取 4 窝不同种系的大白鼠，每窝 4 只，随机分配到 4 个不同的剂量组进行实验。一段时间后，处死大鼠并称量其子宫重量，以比较雌激素剂量和种系的不同水平下大鼠子宫重量之间的差异有无统计学意义。请给出具体的设计方案。

实验前期准备：对大鼠进行编号，第 1 窝编号范围为 1~4，第 2 窝编号范围为 5~8，第 3 窝编号范围为 9~12，第 4 窝编号范围为 13~16。依题意，宜选用随机区组设计。

SAS 程序名为 SASTJFX5_14.SAS。

```
proc plan seed=19830714;
  factors brood=4 ordered dose=4;
  output out=a
  brood cvals=('wobie1' 'wobie2'
               'wobie3' 'wobie4')
  dose cvals=('dose1' 'dose2'
              'dose3' 'dose4');
run;

data b;
  set a;
  mouse=_n_;
run;

proc sort data=b out=c;
  by brood dose;
run;
```

```
proc transpose
  data=c(rename=(dose=_name_))
  out=d(drop=_NAME_);
  var mouse;
  by brood;
run;

ods html;
proc print noobs;
run;
ods html close;
```

输出结果：

| brood | dose1 | dose2 | dose3 | dose4 |
|--------|-------|-------|-------|-------|
| wobie1 | 2 | 4 | 3 | 1 |
| wobie2 | 5 | 7 | 8 | 6 |
| wobie3 | 9 | 10 | 11 | 12 |
| wobie4 | 15 | 14 | 13 | 16 |

结果解释：“brood”列代表大鼠的窝别，其 4 个水平(wobie1、wobie2、wobie3、wobie4)分别代表 4 窝不同种系的大白鼠。dose1、dose2、dose3、dose4 分别代表 4 种不同剂量的雌激素，其下的数字为大鼠的编号。如第 1 行则表示第 1 窝的 4 只大鼠编号为 1~4，按顺序对应分别接受剂量 4、剂量 1、剂量 3、剂量 2 的雌激素处理。

5.2.2 如何用 SAS 实现平衡不完全区组设计

【例 5-15】 研究者欲比较 7 种处理对某观测指标的影响之间的差别有无统计学意义，根据研究对象的属性，将其分为 7 个区组，但由于受试对象较为稀缺，每个区组中仅能保证 3 个受试对象，故研究者准备采用平衡不完全区组设计安排此实验。请给出具体的设计方案。

实验前期准备：对受试对象进行编号，区组 1 内的受试对象编号范围为 1~3，区组 2 内的受试对象编号范围为 4~6，区组 3 内的受试对象编号范围为 7~9，区组 4 内的受试对象编号

范围为 10~12, 区组 5 内的受试对象编号范围为 13~15, 区组 6 内的受试对象编号范围为 16~18, 区组 7 内的受试对象编号范围为 19~21。依题意, 应选用平衡不完全区组设计。

SAS 程序名为 SASTJFX5_15.SAS。

```

%let quzu=7; %let zhiliao=3;

data a;
  do treatment=1 to &quzu;
    output;
  end;
run;

proc optex data=a seed=19830714
coding=orth;
  class Treatment;
  model Treatment;
  blocks
  structure = (&quzu)&zhiliao;
  output out=b;
run;

data c;
  do block=1 to &quzu;
    do treat=1 to &quzu;
      output;
    end; end;
run;

data d;
  set b;
  n1=_n_;
  drop block;
run;

data e;
  set c;
  n2=_n_;
run;

proc sql;
  create table f as
  select * from d,e
  where
  n2=int(n1/(&zhiliao+0.1))
  * &quzu+treatment;
run; quit;

data g;
  merge e f;
  by block treat n2;
  drop n1 n2;
  if treatment^= .
  then treatment=n1;
  drop n1 n2;
run;

proc transpose
  data=g(rename=(treat=_name_))
  out=h(drop=_name_);
  var treatment;
  by block;
run;

data k;
  set h;
  rename
  _1 - _&quzu=treat1 - treat&quzu;
run;

ods html;
proc print noobs;
run;
ods html close;

```

输出结果:

| block | treat1 | treat2 | treat3 | treat4 | treat5 | treat6 | treat7 |
|-------|--------|--------|--------|--------|--------|--------|--------|
| 1 | 1 | . | 3 | . | . | 2 | . |
| 2 | . | 4 | . | . | 6 | 5 | . |
| 3 | . | . | 9 | 7 | 8 | . | . |
| 4 | 10 | 11 | . | 12 | . | . | . |
| 5 | . | . | . | 15 | . | 13 | 14 |
| 6 | 16 | . | . | . | 17 | . | 18 |
| 7 | . | 20 | 19 | . | . | . | 21 |

结果解释: “block”列代表区组, 其 7 个水平(1, 2, …, 7)分别代表 7 个不同的区组。treat1~treat7 分别代表 7 种不同的处理, 其下的数字内容为受试对象的编号。如区组 1 内受试对象 1、3、2 分别接受处理 1、处理 3 和处理 6 的干预。

5.2.3 如何用 SAS 实现拉丁方设计

【例 5-16】 在行走速度和行程固定的前提下，负荷越重，体能消耗越多。目的：研究在 4 种不同的负荷量条件下，消耗的体能之间的差别是否有统计学意义。拟用 4 名战士，在 4 个不同的时期进行实验，并且每人每天只接受 1 种负荷量实验 1 次，假设 4 种负荷量分别为 10kg、20kg、30kg 和 40kg。已知因素之间的交互作用可忽略不计，研究者拟采用拉丁方设计进行此实验。请给出具体的设计方案。

实验前期准备：对 4 名战士进行编号，范围为 soldier1 ~ soldier4。题目中已明确要求选用拉丁方设计。

SAS 程序名为 SASTJFX5_16.SAS。

```
proc plan seed=19830714;
  factors soldier=4 ordered
  day=4 ordered;
  treatments load=4 cyclic;
  output out=a
  soldier cvals=('soldier1'
    'soldier2' 'soldier3'
    'soldier4') random
  day cvals=('day1' 'day2'
    'day3' 'day4') random
  loadcvals=('10kg' '20kg' '30kg'
    '40kg') random;
run;

proc sort data=a out=b;
  by soldier day;
run;
```

```
proc transpose
  data=b(rename=(day=_name_))
  out=c(drop=_name_);
  by soldier;
  var load;
run;

ods html;
proc print data=c noobs;
run;
ods html close;
```

输出结果：

| soldier | day1 | day2 | day3 | day4 |
|----------|------|------|------|------|
| soldier1 | 20kg | 40kg | 10kg | 30kg |
| soldier2 | 10kg | 30kg | 40kg | 20kg |
| soldier3 | 30kg | 10kg | 20kg | 40kg |
| soldier4 | 40kg | 20kg | 30kg | 10kg |

结果解释：“solider”列代表战士，其 4 个水平(soldier1、soldier 2、soldier 3、soldier 4)分别代表 4 名战士。day1、day2、day3、day4 分别代表 4 个不同的时期，其下的内容为负荷量。如第 1 行则表示第 1 名战士 4 个时期内承受的负荷量分别为 20kg、40kg、10kg、30kg。

5.2.4 如何用 SAS 实现 2 × 2 交叉设计

【例 5-17】 某麻醉师研究催醒宁对氟哌啶的作用，选用生理盐水作为催醒宁的对照药，拟用 12 只大鼠作为受试对象，计划采用交叉设计进行此实验。将 12 只大鼠随机分为 2 组，每组 6 只，每只大鼠都要接受催醒宁和生理盐水的处理，其中的 6 只大鼠在第一阶段先接受催醒宁处理，然后在第二阶段接受生理盐水的处理；另外 6 只大鼠在第一阶段先接受生理盐水处理，然后在第二阶段接受催醒宁的处理。请给出具体的设计方案。

实验前期准备：对 12 只大鼠进行编号，范围为 1 ~ 12。对药物进行编号，1 号药物为催醒宁，2 号药物为生理盐水。题目中已明确要求要求进行交叉设计。

SAS 程序名为 SASTJFX5_17.SAS。

```
proc plan seed=19830714;
  factors mouse=12
  time=2 ordered;
  treatments drug=2 cyclic;
  output out=a
  time cvals=('time1' 'time2')
  drug cvals=('drug1' 'drug2');
run;

proc sort data=a out=b;
  by mouse;
run;

proc transpose data=b(rename=
  (time=_name_))
  out=c(drop=_name_);
  by mouse;
  var drug;
run;
```

```
proc sort;
  by time1;
run;

ods html;
proc print data=c noobs;
run;
ods html close;
```

输出结果：

| mouse | time1 | time2 |
|-------|-------|-------|
| 2 | drug1 | drug2 |
| 4 | drug1 | drug2 |
| 6 | drug1 | drug2 |
| 7 | drug1 | drug2 |
| 9 | drug1 | drug2 |
| 10 | drug1 | drug2 |
| 1 | drug2 | drug1 |
| 3 | drug2 | drug1 |
| 5 | drug2 | drug1 |
| 8 | drug2 | drug1 |
| 11 | drug2 | drug1 |
| 12 | drug2 | drug1 |

结果解释：“mouse”列代表大鼠编号，time1 和 time2 分别代表两个实验阶段，time1 列和 time2 列下的内容 (drug1 或 drug2) 分别代表具体的给药情况。drug1 和 drug2 分别代表催醒宁和生理盐水。

5.2.5 如何用 SAS 实现 3×3 交叉设计

【例 5-18】 假定某人为了研究三种药物对高血压的治疗效果，特别希望考察 3 种药物先后作用于同一位患者身上会产生怎样的疗效。现从高血压患者中随机选取 18 名，拟采用 3×3 交叉设计来安排此实验，观察每次服药后的血压值。请给出具体的设计方案。

特别说明：此试验在实际使用时应慎用，因为每个个体身上先后使用了三种不同的药物，而且，用药顺序被随机化了，在多个个体身上观察到每种药物的疗效是随机的、非线性的，但对试验结果采取此种设计下定量资料方差分析处理时，假定其疗效是线性的，故其分析结果很可能是错误的。此处仅仅展示如何用 SAS 呈现此种试验设计类型。若不是三种药物，而是三

台测定体重的仪器，且观测指标是“体重”，则这样设计的结果就非常准确了。以后，遇到类似情况，也应同样理解和对待，不再赘述。

实验前期准备：对 18 名高血压患者进行编号，范围为 1 ~ 18。题目中要求进行 3 × 3 交叉设计。

SAS 程序名为 SASTJFX5_18.SAS。

```
proc plan seed=19830714;
  factorspatient=18 time=3
  ordered;
  treatments drug=3 perm;
  output out=a
  time cvals=('time1' 'time2'
    'time3') random
  drug cvals=('drug1' 'drug2'
    'drug3') random;
run;

proc sort data=a out=b;
  by patient time;
run;

proc transpose
  data=b(rename=(time=_name_))
  out=c(drop=_name_);
  by patient;
  var drug;
run;

proc sort;
  by time1 time2;
run;

ods html;
proc print data=c noobs;
run;
ods html close;
```

输出结果：

| patient | time1 | time2 | time3 |
|---------|-------|-------|-------|
| 6 | drug1 | drug2 | drug3 |
| 7 | drug1 | drug2 | drug3 |
| 14 | drug1 | drug2 | drug3 |
| 3 | drug1 | drug3 | drug2 |
| 4 | drug1 | drug3 | drug2 |
| 15 | drug1 | drug3 | drug2 |
| 5 | drug2 | drug1 | drug3 |
| 12 | drug2 | drug1 | drug3 |
| 17 | drug2 | drug1 | drug3 |
| 2 | drug2 | drug3 | drug1 |
| 8 | drug2 | drug3 | drug1 |
| 11 | drug2 | drug3 | drug1 |
| 1 | drug3 | drug1 | drug2 |
| 13 | drug3 | drug1 | drug2 |
| 16 | drug3 | drug1 | drug2 |
| 9 | drug3 | drug2 | drug1 |
| 10 | drug3 | drug2 | drug1 |
| 18 | drug3 | drug2 | drug1 |

结果解释：“patient”列代表高血压患者编号，time1、time2 和 time3 分别代表三个用药阶段，time1、time2 和 time3 列下的内容（drug1、drug2 或 drug3）分别代表具体的给药情况。drug1、drug2 和 drug3 分别代表三种不同的降血压药物。

5.2.6 如何用 SAS 实现裂区设计

【例 5-19】 某研究者为研究不同瘤株的生瘤效果和不同蛇毒浓度的抑瘤作用，考虑到接种瘤株后需待肿瘤生长一段时间，才进行注射蛇毒的操作，即实验因素“不同瘤株”和“不同浓度的蛇毒”作用于小鼠并非同步，故采用裂区设计来安排此实验。先将 48 只小鼠按原始体重分为 3 个区组，使每个区组中的 16 只小鼠体重最接近，然后将每个区组内的 16 只小鼠再随机均分 4 组，分别接种 4 种不同的瘤株。一段时间后，再随机决定每个区组中接种同一种瘤株的 4 只小鼠分别注射 4 种不同浓度的蛇毒，连续注射蛇毒抑瘤一段时间后处死小鼠，解剖称量瘤重。请给出具体的设计方案。

实验前期准备：对 48 只小鼠进行编号，区组 1 内的小鼠编号范围为 1~12，区组 2 内的小鼠编号范围为 13~24，区组 3 内的小鼠编号范围为 25~36，区组 4 内的小鼠编号范围为 37~48。依题意，应选用裂区设计。

SAS 程序名为 SASTJFX5_19.SAS。

```
%let factor_a=4;

proc plan seed=19830714;
  factors block=3 ordered r=16;
  output out=aa;
run;

data bb;
  set aa;
  geti=_n_;
  do i=1 to &factor_a;
    if r <= i* &factor_a and
       r > (i-1)* &factor_a then do;
      a=i; b=r - (i-1)* &factor_a;
    end; end;
run;

proc sort;
  by block a b;
run;

proc transpose out=cc;
  var a b geti;
  by block;
run;

data dd;
  set cc;
  if _name_='a' or _name_='b' then
    delete;
  drop _name_;
  rename col1 - col4 = a1b1 - a1b4
         col5 - col8 = a2b1 - a2b4
         col9 - col12 = a3b1 - a3b4
         col13 - col16 = a4b1 - a4b4;
run;

ods html;
proc print noobs;
run;
ods html close;
```

输出结果：

| block | a1b1 | a1b2 | a1b3 | a1b4 | a2b1 | a2b2 | a2b3 | a2b4 | a3b1 | a3b2 | a3b3 | a3b4 | a4b1 | a4b2 | a4b3 | a4b4 |
|-------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| 1 | 12 | 13 | 9 | 11 | 16 | 5 | 15 | 4 | 3 | 14 | 10 | 8 | 7 | 1 | 2 | 6 |
| 2 | 21 | 25 | 31 | 18 | 29 | 28 | 17 | 22 | 30 | 26 | 19 | 27 | 32 | 20 | 23 | 24 |
| 3 | 40 | 42 | 47 | 44 | 46 | 43 | 36 | 33 | 48 | 35 | 45 | 37 | 41 | 39 | 38 | 34 |

结果解释：“block”代表小鼠的区组，其 3 个水平(1~3)分别代表 3 个不同的区组。a 代表不同的瘤株，有 4 个水平，即 a1、a2、a3、a4；b 代表不同浓度的蛇毒，有 4 个水平，即 b1、b2、b3、b4。a1b1~a4b4 分别代表这两个因素各水平组合而成的 16 种实验条件，这 16 列下的数字代表接受此种处理的小鼠编号。

实际实验时，是先按区组因素将 48 只小鼠分为 3 个区组，每个区组 16 只小鼠。然后将每个区组中的 16 只小鼠随机等分为 4 小组，每小组 4 只，分别接种 4 种不同的瘤株；接种瘤株成

功后，再随机决定每个区组内 4 个小组中的 4 只小鼠接受 4 种不同浓度的蛇毒处理，一段时间后，称量瘤重。

可见，整个实验分为两个阶段：第一阶段接种瘤株，第二阶段注射蛇毒。所以，在查看上述程序结果时，也应根据这两个阶段进行分组。以第一行为例，这 16 只小鼠来自于区组 1，编号为 1~16。在实验的第一阶段，a1b1~a1b4(均包含 a1)列下的 4 只小鼠(编号为 12、13、9、11)接种瘤株 1；a2b1~a2b4(均包含 a2)列下的 4 只小鼠(编号为 16、5、15、4)接种瘤株 2；a3b1~a3b4(均包含 a3)列下的 4 只小鼠(编号为 3、14、10、8)接种瘤株 3；a4b1~a4b4(均包含 a4)列下的 4 只小鼠(编号为 7、1、2、6)接种瘤株 4。在实验的第二个阶段，每个小组中，实验条件包含 b1 的列下的小鼠接受浓度 1 的蛇毒注射；实验条件包含 b2 的列下的小鼠接受浓度 2 的蛇毒注射；实验条件包含 b3 的列下的小鼠接受浓度 3 的蛇毒注射；实验条件包含 b4 的列下的小鼠接受浓度 4 的蛇毒注射。

5.2.7 如何用 SAS 实现析因设计

【例 5-20】 某医生欲研究回心草各单体成分对实验性心肌缺血血流动力学的影响，研究因素为回心草单体成分(包括胡椒碱、胡椒酸甲酯、咖啡酸甲酯)及其剂量(包括 1nmol/L、10nmol/L、100nmol/L)。现拟采用析因设计进行此实验，选取健康新西兰家兔 36 只，将其随机等分成 9 组，每组 4 只。所有家兔处死后，造缺血缺氧的离体心脏模型，给以各实验组相应单体成分及浓度的药物进行实验，记录各组实验家兔血流动力学的有关指标。请给出具体的设计方案。

实验前期准备：对 36 只新西兰家兔进行编号，范围为 1~36。题目中要求进行析因设计。SAS 程序名为 SASTJFX5_20.SAS。

```
proc plan seed=19830714;
  factors monomer=3 dose=3
  repeat=4;
  output out=a
  monomer cvals=('hujiaojian'
    'hujiaosuan' 'kafeisuan')
  dose nvals=(1 10 100);
run;

proc plan seed=19830714;
  factors subject=36;
  output out=b;
run;

data c;
  merge a b;
run;

proc sort data=c
  out=d(drop=repeat);
  by monomer dose subject;
run;

proc transpose data=d
  out=e(drop=_name_);
  var subject;
  by monomer dose;
run;

data h;
  set e;
  rename col1-col4=
  repeat1-repeat4;
run;

ods html;
proc print noobs;
run;
ods html close;
```

输出结果：

| monomer | dose | repeat1 | repeat2 | repeat3 | repeat4 |
|------------|------|---------|---------|---------|---------|
| hujiaojian | 1 | 5 | 6 | 23 | 31 |
| hujiaojian | 10 | 7 | 20 | 28 | 29 |

| | | | | | |
|------------|-----|----|----|----|----|
| hujiaojian | 100 | 2 | 8 | 14 | 19 |
| hujiaosuan | 1 | 10 | 12 | 25 | 26 |
| hujiaosuan | 10 | 3 | 4 | 35 | 36 |
| hujiaosuan | 100 | 1 | 13 | 21 | 33 |
| kafeisuan | 1 | 9 | 22 | 27 | 34 |
| kafeisuan | 10 | 11 | 16 | 17 | 24 |
| kafeisuan | 100 | 15 | 18 | 30 | 32 |

结果解释：“monomer”列代表回心草单体成分，其3个水平(hujiaojian、hujiaosuan 和 kafeisuan)分别为胡椒碱、胡椒酸甲酯、咖啡酸甲酯。“dose”列代表各单体所用剂量，其3个水平(1、10 和 100)分别代表 1nmol/L、10nmol/L、100nmol/L。repeat1、repeat2、repeat3、repeat4 分别代表每种实验条件下的4次独立重复实验，这4列下的数据即为进行独立重复实验的新西兰家兔的编号。如第一行表示编号为5、6、23、31的4只新西兰家兔接受剂量为1nmol/L的胡椒碱处理。

5.2.8 如何用 SAS 实现含区组因素的析因设计

【例 5-21】 某研究者欲研究小鼠在以不同注射剂量(每天总注射量，因素 A)和不同注射频率(每天注射次数，因素 B)下药剂 ACTH 对尿总酸度的影响。其中，A 有 3 个水平：A1、A2、A3(即 3 种不同的日注射量)；B 有 3 个水平：B1、B2、B3(3 种不同的日注射次数)。这两个因素共有 9 种水平组合，称为 9 个处理组。研究者欲采用含区组因素的析因设计安排此实验，先将 54 只小鼠按某些条件分为 6 个区组，使每个区组中的 9 只小鼠在规定的方面条件最为接近，然后，将每个区组中的 9 只小鼠随机地分入 9 个处理组中去。请给出具体的设计方案。

实验前期准备：对 54 只小鼠进行编号，区组 1 内的小鼠编号范围为 1~9，区组 2 内的小鼠编号范围为 10~18，区组 3 内的小鼠编号范围为 19~27，区组 4 内的小鼠编号范围为 28~36，区组 5 内的小鼠编号范围为 37~45，区组 6 内的小鼠编号范围为 46~54。题目中要求进行含区组因素的析因设计。

SAS 程序名为 SASTJFX5_21.SAS。

```

%let factor_a=3;
proc plan seed=19830714;
  factors block=6 ordered r=9;
  output out=aa;
run;

data bb;
  set aa;
  geti=_n_;
  do i=1 to &factor_a;
    if r <= i* &factor_a and
       r > (i-1)* &factor_a then do;
      a=i; b=r - (i-1)* &factor_a;
    end; end;
run;

proc sort;
  by block a b;
run;

proc transpose out=cc;
  var a b geti;
  by block;
run;

data dd;
  set cc;
  if _name_='a' or _name_='b'
  then delete;
  drop _name_;
  rename col1-col3=a1b1-a1b3
         col4-col6=a2b1-a2b3
         col7-col9=a3b1-a3b3;
run;

ods html;
proc print noobs;
run;
ods html close;

```

输出结果：

| block | a1b1 | a1b2 | a1b3 | a2b1 | a2b2 | a2b3 | a3b1 | a3b2 | a3b3 |
|-------|------|------|------|------|------|------|------|------|------|
| 1 | 2 | 8 | 4 | 7 | 5 | 3 | 9 | 1 | 6 |
| 2 | 11 | 17 | 10 | 15 | 14 | 12 | 13 | 16 | 18 |
| 3 | 27 | 25 | 24 | 21 | 22 | 19 | 26 | 20 | 23 |
| 4 | 34 | 33 | 36 | 32 | 30 | 35 | 29 | 31 | 28 |
| 5 | 43 | 39 | 45 | 40 | 44 | 37 | 41 | 38 | 42 |
| 6 | 48 | 49 | 46 | 52 | 54 | 53 | 51 | 47 | 50 |

结果解释：“block”代表小鼠的区组，其6个水平(1~6)分别代表6个不同的区组。a代表不同的注射剂量，有3个水平，即a1、a2、a3；b代表不同的注射频率，有3个水平，即b1、b2、b3。a1b1~a3b3分别代表这两个因素各水平组合而成的9种实验条件，这9列下的数字代表接受此种处理的小鼠编号。

5.3 本章小结

本章将常用的各种多因素设计定量资料以便于理解且比较规范化的形式呈现出来，有利于使用者辨析实验设计类型，有利于统计分析方法的合理选择。同时，本章介绍了一些常用多因素设计类型的 SAS 实现，有利于科研工作者套用本书内程序进行相应的试验分组安排。

(高辉 胡完)

第6章 样本含量与检验效能估计的 SAS 实现

目前,不管是从期刊上发表的论文,还是从实际科研人员的实验记录等科研档案来看,绝大多数研究工作都没有交代样本含量的确定方法、计算依据及公式,没有进行科学的、有理有据的样本含量估计。当假设检验的结果为“阴性”,即不能拒绝原假设时,也很少有研究者去分析假设检验的检验效能,进而判断“阴性”结果的可信度如何。造成这种状况的原因:一是研究者没有认识到样本含量估计与检验效能分析对科学研究的重要性,二是研究者不知道如何进行样本含量估计与检验效能分析。为此,本章将为读者介绍常见情况下样本含量估计与检验效能分析的具体方法,并给出相应的 SAS 程序。

6.1 估计样本含量与检验效能的前提条件

在进行统计学的教学、咨询和培训工作中,统计学工作者经常被问到的一个问题是:“ $\times \times$ 教授(老师),我希望说明一个新的治疗方案优于旧的治疗方案,请您告诉我该选用多少名患者?”这个问题常使统计学工作者无言以对,因为要想确定一个研究项目所需要的样本含量,凭空想象是没有科学依据的,必须提供一些起码的前提条件,才能做出有理有据的估算。估计样本含量时,通常需要提供的前提条件有如下几项。

(1)定出检验水准或显著性水平:事先规定本次试验允许犯 I 型(或假阳性)错误的概率 α ,通常规定 $\alpha = 0.05$,同时还应明确是单侧检验还是双侧检验。 α 定得越小,所需的样本含量越大。同一问题,在其他条件不变的情况下,用单侧检验比用双侧检验所需的样本含量要少,但究竟该用单侧检验还是双侧检验取决于专业知识,而不是人的意愿。

(2)提出所期望的检验效能或把握度 $\text{power} = 1 - \beta$ (这里的 β 为允许犯 II 型或假阴性错误的概率),即在特定的 α 水准下,若总体对比的参数之间确实存在着差别,此时该次试验能发现此差别的概率。要求的检验效能越大,所需的样本含量就越大,实际上,检验效能由犯 II 型错误的概率 β 的大小所决定。在科研设计时,检验效能不宜低于 0.75,否则检验的结果很可能反映不出总体的真实差别,出现非真实的阴性结果。

(3)必须知道由样本推断总体的一些信息。比较两总体均数或概率的差别时,应当知道总体参数间的差值 δ 的信息。比如两总体均数间的差值 $\delta = \mu_1 - \mu_2$ 的信息,两总体概率间的差值 $\delta = \pi_1 - \pi_2$ 的信息。有时,研究者很难得到总体参数的信息,但可以用专业上(临床上)认为有意义的差值代替,如平均舒张期血压的差值 $\geq 0.69 \text{ kPa} (5 \text{ mmHg})$,白细胞的平均差值为 $0.5 \times 10^9 / \text{L} (500 \text{ 个} / \text{mm}^3)$ 等,也有人主张用 0.25 倍或 0.50 倍的标准差估计总体均数间的差值。当然,也可以根据试验的目的人为规定,如规定试验的新药有效概率必须超过标准药物有效概率的 30% 才有推广价值等。此外,确定两均数比较的样本含量还需要有关总体标准差 σ 的信息。这些信息也可以通过查阅资料,借鉴前人的经验或者进行预试验来寻找参考值。

6.2 抽样调查中样本含量估计

为了保证调查研究的质量,必须保证足够的观测单位数量,这就是样本含量的问题。样本含量的大小与抽样误差有直接关系,因此确定一个恰当的样本含量可以将抽样误差控制在可接受的范围。

6.2.1 估计总体均值时如何估计样本含量

【例6-1】 某临床医师欲了解高脂血症患者的血液中甘油三酯的含量,现已知正常人血液中甘油三酯含量的标准差是0.35,这次希望绝对误差不超过0.2,取 $\alpha = 0.05$,在以上条件下,要估计患者的平均甘油三酯含量,需要调查多少患者?

【分析与解答】 根据题目中的信息,已知总体标准差 $\sigma = 0.35$ 、允许误差 $\Delta = 0.2$,以上计算可以通过 SAS 程序完成,程序名为 SASTJFX6_1.SAS。

```
proc power;
  onesamplemeans ci=t sides=2 alpha=0.05 halfwidth=0.2 stddev=0.35
  ntotal=. probwidth=0.3;
run;
quit;
```

运行结果:需调查大约 12 人。

【程序说明】 选项“ci=t”表示要估计总体均数的基于 t 分布的置信区间,该语句也可以简写为“ci”;选项“sides=2”(默认设置)中的 2 表示要求双侧对称的置信区间;需要注意的是,选择双侧还是单侧,不是随意确定的,而是以专业知识为依据,如果能够确定该指标不会高于或低于某一个值,则选择单侧。选项“alpha=0.05”(默认设置)中的 0.05 表示要求置信度为 100(1 - 0.05)% 的置信区间;选项“halfwidth=0.2”中的 0.2 为误差,即置信区间的半宽度;选项“stddev=0.35”中的 0.35 为已知的标准差;选项“probwidth=0.30”中的 0.30 为获得半宽不超过最大误差值 0.2 的置信区间的概率,此值定得越大,所需要的样本量也会越大,如取“probwidth=0.90”,则 n=20。在应用时,读者只要根据实际情况修改这些关键字所代表的变量的取值,即可得到结果。在实际使用中,为保险起见,“probwidth”的值应取大一些,如 0.95 或 0.99。

6.2.2 估计总体率时如何估计样本含量

1. 当 π 或 P 接近 0 或 1 时

【例6-2】 某地区现调查 HBsAg 阳性率,过去调查的结果为 P=10%,本次调查容许误差不超过 0.1P, $\alpha = 0.05$ (双侧),估计应调查人数。

【分析与解答】 该研究是一个结果变量为二值变量的单组设计定性资料阳性率估计问题,已知阳性率的估计值、阳性率的调查容许误差(即阳性率置信区间的半宽)及显著性水平,欲估计所需调查人数。由于 P=10%,比较接近于 0,故需按下面的 SAS 程序进行样本含量估计(当 P 比较接近于 1 时也应如此),程序名为 SASTJFX6_2.SAS。

```
%let P=0.10; %let delta=0.01;
%let alpha=0.05;
data d6_2;
  n = (probit(1 - &alpha/2) /
  arsin(&delta/sqrt(&P*(1 - &P))))**2;
  n1 = int(n) + 1;
  file print;
  PUT #3 @15 '需调查大约' n1 '人.';
run;quit;
```

运行结果：需调查大约 3457 人。

【程序说明】第 1 行中 $P=0.10$ 中的 0.10 为阳性率， $\delta=0.01$ 中的 0.01 是误差的限值， $\alpha=0.05$ 中的 0.05 为可能犯 I 型错误的概率，上述数字需要根据实际情况修改。

2. 当 π 或 P 接近 0.5 时

【例 6-3】拟用抽样调查了解某地小学生蛔虫感染率。假定以往该地小学生蛔虫感染率 $P=50\%$ ，要求误差不超过 3%，如取 $\alpha=0.05$ 。问：需调查多少人？

【分析与解答】该研究是一个结果变量为二值变量的单组设计定性资料感染率估计问题，已知感染率的估计值、感染率的调查容许误差（即感染率置信区间的半宽）及显著性水平，欲估计所需调查人数。由于 $P=50\%$ ，故需按下面的 SAS 程序进行样本含量估计（ P 比较接近于 50% 时均应如此；当 P 未知时，按 $P=0.50$ 计算），程序名为 SASTJFX6_3.SAS。

```
%let p=0.5; %let delta=0.03;
%let alpha=0.05;
data d6_3;
n = (probit((1 - &alpha/2))/&delta);
** 2* &p* (1 - &p);
n1 = int(n) + 1;
```

```
file print;
PUT #3 @15 '需调查大约' n1 '人。';
run;quit;
```

运行结果：需调查大约 1068 人。

【程序说明】第 1 行中 $P=0.5$ 中的 0.5 为感染率， $\delta=0.03$ 中的 0.03 是误差的限值， $\alpha=0.05$ 中的 0.05 为可能犯 I 型错误的概率，上述数字需要根据实际情况修改。

6.3 定量资料假设检验中样本含量与检验效能估计

6.3.1 单组、配对或交叉设计定量资料统计分析时样本含量估计

【例 6-4】用某药治疗矽肺患者后，尿矽排出量平均增加 15mg/L，其标准差为 12mg/L。假定该药确能使尿矽排除量增加，定 $\alpha=0.05$ （单侧）， $\beta=0.10$ 。问：需观察多少患者才能得出服药前后尿矽排除量之间差别有统计学意义的结论？

【分析与解答】本研究，若以尿矽排除量作为结果变量，则该研究资料所对应的实验设计类型为自身配对设计；若以尿矽排除量增加量作为结果变量，则该研究资料所对应的实验设计类型为单组设计，但不管是配对设计还是单组设计，本质上都是单因素 2 水平问题，对应的分析方法本质上都是一样的。事实上，配对设计的资料往往都是转化为单组设计资料进行处理的。这里，已知样本均值、样本标准差、单侧检验的显著性水平及容许犯 II 型错误的概率，所要达到的目的不是对资料进行统计分析，而是进行样本含量估计。所需的 SAS 程序如下。程序名为 SASTJFX6_4.SAS。

```
proc power;
onesamplemeans test = t
sides = 1
mean = 15
stddev = 12
```

```
ntotal = .
power = 0.90;
run;quit;
```

运行结果：需观察 8 例患者。

【程序说明】程序中的选项“test = t”表示对均值进行 t 检验；sides = 1 中的 1 表示进行单侧检验，若 sides = 2 或者省略则表示进行双侧检验；mean = 15 中的 15 为总体均值的估计值；stddev = 12 中的 12 为总体标准差的估计值；ntotal = .（缺失值）表示要估计总的样本含量；power = 0.90 中的 0.90 为检验效能的值。程序中的数字需要根据实际情况进行修改。

6.3.2 成组设计统计分析时样本含量估计

【例6-5】 在动物镇咳试验中,比较中药复方 I 与复方 II 使小鼠推迟发生咳嗽的时间,复方 I 与复方 II 的平均数分别为 31.67s 和 44.00s (即 $\delta = 44.00 - 31.67 = 12.33$ s)。设两组标准差相等,且为 25s, $\alpha = 0.05$ (双侧), $\beta = 0.10$, 要得出两组之间的差别有统计学意义的结论,需要用多少只小鼠?

【分析与解答】 这是一个结果变量为定量变量的成组设计研究的样本含量估计问题。已知两组的样本均值、样本标准差及假设检验中容许犯 I、II 型错误的概率,假定两组的样本含量相等,欲估计最低样本含量。所需的 SAS 程序如下。程序名为 SASTJFX6_5.SAS。

```
proc power;
  twosamplemeans
    test=diff
    sides=2
    meandiff=12.33
    stddev=25.00
    groupweights=(1 1)
    ntotal=.
    power=0.90;
run;quit;
```

运行结果: 两组共需 176 只小鼠, 每组需要 88 只小鼠。

【程序说明】 程序中的选项“test = diff”表示对均差进行 t 检验; “meandiff = 12.33”中的 12.33 为两均值的差值; “groupweights = (1 1)”表示两组样本含量的分配比为 $n_1:n_2 = 1:1$; 其他各选项的说明参见例 6-4。程序中的数字需要根据实际情况修改。

6.3.3 成组设计等效性检验时样本含量估计

【例6-6】 某利尿新药拟进行 II 期临床试验,与阳性药按 1:1 的比例安排例数,考察 24h 新药利尿量(mL)是否相当于阳性药。根据以往的疗效和统计学的一般要求,取 $\alpha = 0.05$, $\beta = 0.20$, 等效标准 $\delta = 60$ mL, 新药与阳性对照药利尿量的差值 θ 估计为 -20 mL, 已知两组共同标准差 $s_c = 180$ mL。问: 每组需要多少病例?

【分析与解答】 由研究目的可知,这是一个成组设计定量资料的等效性试验样本含量估计问题,已知试验组与对照组的样本含量比例、两组利尿量差值的估计值、共同的样本标准差、等效界值及假设检验中容许犯 I、II 型错误的概率,欲估计各组的最低样本含量。所需的 SAS 程序如下,程序名为 SASTJFX6_6.SAS。

```
%let alpha=0.05;
%let beta=0.20;
%let s=180;
%let delta=60;
%let theta=-20;
%let k=1;
%let max_n=1000;
data a;
do n_t=2 to &max_n;
p=probnorm(sqrt((&theta-&delta)
** 2* &k* n_t/(&k+1)/&s** 2))-
probit(1-&alpha/2))+
probnorm(sqrt((&theta+&delta)*
** 2* &k* n_t/(&k+1)/&s** 2))-
probit(1-&alpha/2))-1;
if p>=1-&beta then goto ok;
end;
ok:
n_c=n_t* &k;
file print;
PUT #2 @15 '试验组所需要的样本例数为
' n_t '例。';
PUT #3 @15 '对照组所需要的样本例数为
' n_c '例。';
run;
```

运行结果:

试验组所需要的样本例数为 318 例。

对照组所需要的样本例数为 318 例。

【程序说明】 第 1~7 行分别设置可能犯 I 型错误的概率、可能犯 II 型错误的概率、两组的混合标准差、等效性界限值、两组总体均数差值的估计值、对照组与试验组的样本含量比值、最大迭代次数, data 步循环计算试验组各样本含量取值时能够达到的检验效能。若检验效能足够大, 则程序的运行跳出循环, 并打印输出结果。上述数字需要根据实际情况修改。

6.3.4 成组设计非劣效或优效性检验时样本含量估计

【例 6-7】 某利尿新药拟进行 II 期临床试验, 与阳性药按 1:1 的比例安排例数, 考察 24h 新药利尿量(mL)是否不差于阳性药。根据以往的疗效和统计学的一般要求, 取 $\alpha = 0.05$, $\beta = 0.20$, 非劣效界值 $\delta = -60\text{mL}$, 已知两组共同标准差 $s = 180\text{mL}$, 假定新药与阳性对照药总体利尿量的差值 $\theta = -20\text{mL}$ 。问: 每组需要多少病例?

【分析与解答】 由研究目的可知, 这是一个成组设计定量资料的非劣效试验样本含量估计问题。已知试验组与对照组的样本含量比例、两组利尿量差值的估计值、共同的样本标准差、非劣效界值及假设检验中容许犯 I、II 型错误的概率, 欲估计各组的最低样本含量。所需的 SAS 程序如下, 程序名为 SASTJFX6_7.SAS。

```
%let alpha=0.05;
%let beta=0.20;
%let s=180;
%let delta_L=-60;
%let theta=-20;
data a;
n=2*(probit(1-&alpha)+probit(1-
&beta))*2*&s**2/(&delta_L-
&theta)**2;
file print;
PUT #2 @15 '每组所需要的样本例数为'n'例。';
run;
```

运行结果: 每组所需要的样本例数为 250.3935679 例。取 $n = 251$ 。

【程序说明】 第 1~5 行分别设置可能犯 I 型错误的概率、可能犯 II 型错误的概率、两组的混合标准差、非劣效界值和两组总体均数差值的估计值, data 步中则按照算法计算样本含量的估计值, 并打印输出结果。上述数字需要根据实际情况修改。

6.3.5 单因素多水平设计定量资料方差分析时样本含量的估计

【例 6-8】 用三种方法治疗脑卒中抑郁患者, 观察神经功能康复状况。估计治疗后三种方法 SSS 评分均数分别为 11.0, 23.0, 9.0, 标准差分别为 3, 3, 2。如果要得到三组之间的差别有统计学意义的结论, 每组各需要多少例患者?

【分析与解答】 这是一个结果变量为定量变量的单因素 3 水平设计试验研究的样本含量估计问题。所需的 SAS 程序如下。程序名为 SASTJFX6_8.SAS。

```
data d6_8;
input method$ score CellWgt;
datalines;
A 11.0 1
B 23.0 1
C 9.0 1
;
contrast "A vs. B" method 1 -1 0;
contrast "A vs. C" method 1 0 -1;
contrast "B vs. C" method 0 1 -1;
power
stddev=3.0
alpha=0.05
ntotal=.
```

```
run;
proc glmpower data = d6_8;
  class method;
  model score = method;
  weight CellWgt;
run;quit;
```

运行结果:

| The GLMPOWER Procedure | | | | | | |
|--------------------------|----------|---------|---------|----------|--------------|---------|
| Fixed Scenario Elements | | | | | | |
| Dependent Variable | | | score | | | |
| Weight Variable | | | CellWgt | | | |
| Alpha | | | 0.05 | | | |
| Error Standard Deviation | | | 3 | | | |
| Nominal Power | | | 0.9 | | | |
| Computed N Total | | | | | | |
| Index | Type | Source | Test DF | Error DF | Actual Power | N Total |
| 1 | Effect | method | 2 | 6 | 0.989 | 9 |
| 2 | Contrast | A vs. B | 1 | 6 | 0.981 | 9 |
| 3 | Contrast | A vs. C | 1 | 141 | 0.900 | 144 |
| 4 | Contrast | B vs. C | 1 | 6 | 0.997 | 9 |

由计算结果可知：若要求通过单因素 3 水平设计定量资料的方差分析及两两比较有 90% 的把握发现三个组 SSS 评分的总体均值不完全相等，则每组只需 3 例患者，三组共需 9 例患者；但若要求通过单因素 3 水平设计定量资料的方差分析及两两比较有 90% 的把握发现三个组中任何两个组的 SSS 评分的总体均值不相等，则每组需要 48 例患者，三组共需 144 例患者。

【程序说明】 data 步定义影响因素各水平组指标的估计值及各水平组的样本含量分配比例；glmpower 过程中的选项“stddev = 3.0”中的 3.0 为各水平组指标的估计值，本例取三个水平组总体标准差估计值的最大值带入计算，也可以同时填入多个总体标准差的估计值，将选项写成“stddev = 2.0 2.5 3.0”或“stddev = 2.0 to 3.0 by 0.5”的形式，SAS 系统将分别计算出每种标准差情况下对应的所需样本量；其他各选项的说明参见前面各例。程序中的数字需要根据实际情况修改。

6.3.6 两因素析因设计定量资料方差分析时样本含量估计

【例 6-9】 某医院用中药复方治疗高胆固醇血症，拟采用 2 × 2 析因设计方案进行研究。先进行预试验，将 12 例高胆固醇患者随机分为四组，用不同疗法治疗：第一组用一般疗法，第二组在一般疗法上外加用甲药，第三组在一般疗法上外加用乙药，第四组在一般疗法上外加用甲药和乙药，一个月后观察胆固醇降低数(mg%)，资料如下。拟对该资料采用 2 × 2 析因设计定量资料的方差分析处理，要求检验效能达到 80%，试估计样本含量。

- 第 1 组：16, 25, 18
- 第 2 组：56, 44, 42
- 第 3 组：28, 31, 23
- 第 4 组：64, 78, 80。

第一步：先由预试验资料估计各试验条件下胆固醇降低数(mg%)的总体均值和总体标准差。程序如下，程序名为：SASTJFX6_9_1.SAS。


```

data step1;
do A=1 to 2;
do B=1 to 2;
do i=1 to 3;
input x @@ ; output;
end;end;end;
cards;
16 25 18
28 31 23
;
run;
proc glm;
class A B;
model x=A B A*B/ss3;
means A*B;
run; quit;

```

主要输出结果:

| Level of A | Level of B | N | x | |
|---------------|---------------|---|------------|------------|
| | | | Mean | Std Dev |
| 1 | 1 | 3 | 19.6666667 | 4.72581563 |
| 1 | 2 | 3 | 27.3333333 | 4.04145188 |
| 2 | 1 | 3 | 47.3333333 | 7.57187779 |
| 2 | 2 | 3 | 74.0000000 | 8.71779789 |

第二步: 根据上面计算得到的有关估计值进行样本含量估计。程序如下, 程序名为: SAS-TJFX6_9_2.SAS。

```

data step2;
do A=1 to 2;
do B=1 TO 2;
input x @@ ;
output;
end;end;
datalines;
19.7 27.3
47.3 74.0
;
run;

proc glmpower data=step2;
class A B;
model x=A|B;
power
stddev=4.0 to 9.0 by 1.0
ntotal=.
power=0.80;
plot x=power min=0.50 max=0.95
key=byfeature(pos=inset);
run;quit;

```

输出结果如下:

The GLMPower Procedure

Fixed Scenario Elements

Dependent Variable x

Nominal Power 0.8

Alpha 0.05

Computed N Total

| Index | Source | Std Dev | Test DF | Error DF | Actual Power | N Total |
|-------|--------|---------|---------|----------|--------------|---------|
| 1 | A | 4 | 1 | 4 | >.999 | 8 |
| 2 | A | 5 | 1 | 4 | >.999 | 8 |
| 3 | A | 6 | 1 | 4 | >.999 | 8 |
| 4 | A | 7 | 1 | 4 | >.999 | 8 |
| 5 | A | 8 | 1 | 4 | 0.997 | 8 |
| 6 | A | 9 | 1 | 4 | 0.988 | 8 |
| 7 | B | 4 | 1 | 4 | 0.992 | 8 |
| 8 | B | 5 | 1 | 4 | 0.944 | 8 |
| 9 | B | 6 | 1 | 4 | 0.850 | 8 |

| | | | | | | |
|----|-------|---|---|----|-------|----|
| 10 | B | 7 | 1 | 8 | 0.959 | 12 |
| 11 | B | 8 | 1 | 8 | 0.900 | 12 |
| 12 | B | 9 | 1 | 8 | 0.823 | 12 |
| 13 | A * B | 4 | 1 | 8 | 0.950 | 12 |
| 14 | A * B | 5 | 1 | 8 | 0.825 | 12 |
| 15 | A * B | 6 | 1 | 12 | 0.832 | 16 |
| 16 | A * B | 7 | 1 | 16 | 0.817 | 20 |
| 17 | A * B | 8 | 1 | 24 | 0.858 | 28 |
| 18 | A * B | 9 | 1 | 28 | 0.826 | 32 |

例 6-9 检验效能与样本含量的关系曲线如图 6-1 所示。

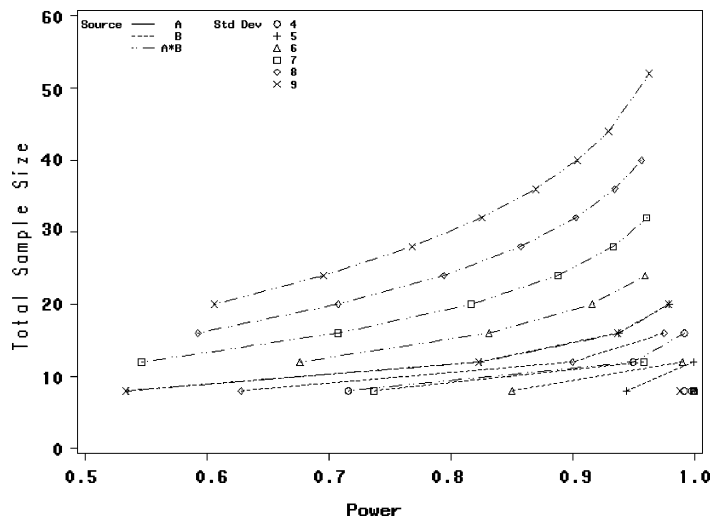


图 6-1 例 6-9 检验效能与样本含量的关系曲线

由输出结果可知：总样本含量达到 8 例就能有 98.8% 以上的把握判断甲药有无疗效，总样本含量达到 12 例就能有 82.3% 以上的把握判断乙药有无疗效，总样本含量达到 32 例就能有 82.6% 以上的把握判断“甲药用否”和“乙药用否”两个因素有无交互效应。

【程序说明】 程序的 data 步用于定义各试验条件下指标的总体均值；glmpower 过程用于功效分析，其中的语句“model x = A|B;”等价于“model x = A B A * B;”，选项“stddev = 4.0 to 9.0 by 1.0”用于定义指标的总体标准差，可以同时列出多个总体标准差。上述数字需要根据实际情况修改。

6.3.7 简单直线相关或回归分析时样本含量的估计

【例 6-10】 估计总体相关系数 $\rho = 0.70$ ，规定 $\alpha = 0.05$ （双侧）， $\beta = 0.10$ 。问：需要多少例？所需的 SAS 程序如下，程序名为：SASTJFX6_10.SAS。

| | |
|--|--|
| <pre>proc power; onecorr dist=t /* npartialvars=0 */ corr=0.70 ntotal=. power=0.90; run;quit;</pre> | <pre>proc power; onecorr dist=fisherz /* npartialvars=0 */ corr=0.70 ntotal=. power=0.90; run;quit;</pre> |
|--|--|

运行结果：所需要的样本含量为 16 例(t)或 17 例(Fisher's z)。

【程序修改指导】 选项“`dist =`”定义对相关系数的假设检验方法；被注释掉的选项“`npartialvars = 0`”用于说明除了程序里的相关系数所直接涉及的 2 个主要的定量变量之外，资料还涉及的其他定量变量的个数(主要用于多重线性回归及偏相关分析场合。因本例为简单直线相关，故取值为 0)。程序中的数字需要根据实际情况修改。

6.3.8 单组、配对或交叉设计定量资料假设检验时检验效能的计算

【例 6-11】 用某药治疗 12 例矽肺患者后，尿矽排除量平均增加 25mg/L，其标准差为 22mg/L。已知该药确能使尿矽排除量增加，取 $\alpha = 0.05$ (单侧)，拟对资料进行配对设计定量资料的 t 检验。请分析其检验效能。

【分析与解答】 这是一个配对设计定量资料假设检验的检验效能分析问题。根据已知条件，可用下面的 SAS 程序进行检验效能分析，程序名为 SASTJFX6_11.SAS。

```
proc power;
onesamplemeans test = t
  mean = 25
  stddev = 22
  nttotal = 12
  sides = 1
  power = .;
run;quit;
```

运行结果：检验效能(Power) = 0.979，检验效能是足够的。

【程序说明】 程序第 2~5 行的各选项依次说明假设检验方法、(差值的)均值、标准差和样本含量(配对设计时样本含量为对子数)，`sides = 1` 表示进行单侧检验。上述数字需要根据实际情况进行修改。

6.3.9 成组设计均值差异性检验时检验效能的计算

【例 6-12】 对某新降压药(试验药)与某老降压药(对照药)进行临床试验，用药后两组病人舒张压(kPa)平均下降情况见表 6-1。试计算此统计分析的检验效能。

表 6-1 两组病人用药后舒张压(kPa)平均下降值

| 药物种类 | 例数 | 舒张压(kPa)下降值 | | P |
|------|----|-------------|-------|-------|
| | | \bar{x} | S_d | |
| 试验药 | 30 | 1.50 | 0.83 | >0.05 |
| 对照药 | 30 | 1.20 | 0.83 | |

【分析与解答】 这是一个成组设计定量资料假设检验的检验效能分析问题。根据已知条件，可用下面的 SAS 程序进行检验效能分析，程序名为 SASTJFX6_12.SAS。

```
proc power;
twosamplemeans
test = diff
groupmeans = 1.50|1.20
stddev = 0.83
groupns = (30 30)
power = .;
run;quit;
```

运行结果：检验效能(Power) = 0.280 < 0.75，检验效能不够充分。

【程序说明】 选项“`groupmeans = 1.50|1.20`”中的 1.50 和 1.20 分别为两组的均数，选项“`stddev = 0.83`”中的 0.83 是两组的混合标准差，选项“`groupns = (30 30)`”中的 30 和 30 分别为两组的例数。上述数字需要根据实际情况修改。

6.3.10 成组设计均值等效性检验时检验效能的计算

【例 6-13】 某试验欲显示一种新药血管紧张素 II 拮抗剂(AII antagonist)与标准药血管紧

张素转换酶抑制剂(ACEI)对于治疗轻中度原发性高血压是否具有等效性。该试验的主要终点指标是仰卧舒张压(SDBP, 单位为 kPa), 以药物 ACE 作为阳性对照, 每组所用病例为 150 例。已知两组合并标准差 $s = 1.06\text{kPa}$, 取等效界值 $\delta = 0.40\text{kPa}$, $\alpha = 0.05$, 并假定 AII 与 ACEI 的降压作用的差值 $\theta = 0.10\text{kPa}$, 试分析检验的效能。

【分析与解答】 这是一个成组设计定量资料的等效性检验效能分析问题。根据已知条件, 可用下面的 SAS 程序进行检验效能分析, 程序名为 SASTJFX6_13. SAS。

```
%let n=150;
%let delta=0.40;
%let theta=0.10;
%let s=1.06;
%let alpha=0.05;
%let k=1;
Data a;

power = probnorm (sqrt ((&theta -
&delta)
** 2* &n/(&k+1)/&s** 2) -probit(1
- &alpha/2 )) + probnorm (sqrt
((&theta +&delta)** 2* &n/
(&k+1)/&s** 2) -probit(1 - &alpha/
2)) -1;
File print;
Put 'power=' power;
Run; quit;
```

运行结果: power=0.6715141804, 即检验效能约为 0.67。

【程序说明】 参见例 6-6。

6.3.11 成组设计均值非劣效或优效性检验时检验效能的计算

【例 6-14】 对例 6-13, 若按非劣性设计试验, 考察 AII 是否不比 ACEI 的疗效差。其他条件不变, 试分析检验的效能。

【分析与解答】 这是一个成组设计定量资料的非劣效性检验效能分析问题。根据已知条件, 可用下面的 SAS 程序进行检验效能分析, 程序名为 SASTJFX6_14. SAS。

```
%let alpha=0.05;
%let delta_L= -0.40;
%let theta=0.10;
%let s=1.06;
%let n=150;
data a;

Power = probnorm (abs (&delta_L -
&theta)
* (2* &s** 2/&n)** (-1/2) -
probit(1 - &alpha));
file print;
put #3 @15 "Power =" Power ;
run;quit;
```

运行结果: 此非劣性检验所能达到的检验效能为 0.993, 即 99.3%。

【程序说明】 程序中的数字需要根据实际情况修改。对于类似的实际问题, 只需修改 5 个“%let”语句中的参数就可以了。

6.3.12 单因素多水平设计定量资料的方差分析时检验效能的计算

【例 6-15】 研究轻度和重度再障贫血患者血清中可溶性 CD₈ 抗原水平 (U/mL) 与正常人之间的差别有无统计学意义, 以反映患者免疫状态紊乱而导致造血功能障碍的程度。从 3 种人群中各随机抽取 10 人, 测得 CD₈ 抗原水平 ($\bar{x} \pm s$) 如下, 并采用单因素 3 水平设计定量资料的方差分析对资料进行了处理。试计算其检验效能。

正常组: 288.9 ± 173.6 。

轻度组: 657.9 ± 154.7 。

重度组：762.9 ± 127.1。

所需的 SAS 程序如下，程序名为：SASTJFX6_15.SAS。

```
proc power;
onewayanova
groupmeans = 288.9 | 657.9 | 762.9
stddev = 173.6 154.7 127.1
groupweights = (1 1 1)
alpha = 0.05
ntotal = 30
power = .;
plot x = n min = 6 max = 45 key = bycurve (pos = inset numbers = off);
run;quit;
```

输出结果如下：

| | | | |
|----------------------------------|---------|--------|-------|
| The POWER Procedure | | | |
| Overall F Test for One-Way ANOVA | | | |
| Fixed Scenario Elements | | | |
| Method | Exact | | |
| Alpha | 0.05 | | |
| Group Means | 288.9 | 657.9 | 762.9 |
| Group Weights | 1 | 1 | 1 |
| Total Sample Size | 30 | | |
| Computed Power | | | |
| Index | Std Dev | Power | |
| 1 | 174 | > .999 | |
| 2 | 155 | > .999 | |
| 3 | 127 | > .999 | |

例 6-15 总样本含量与检验功效的关系如图 6-2 所示。

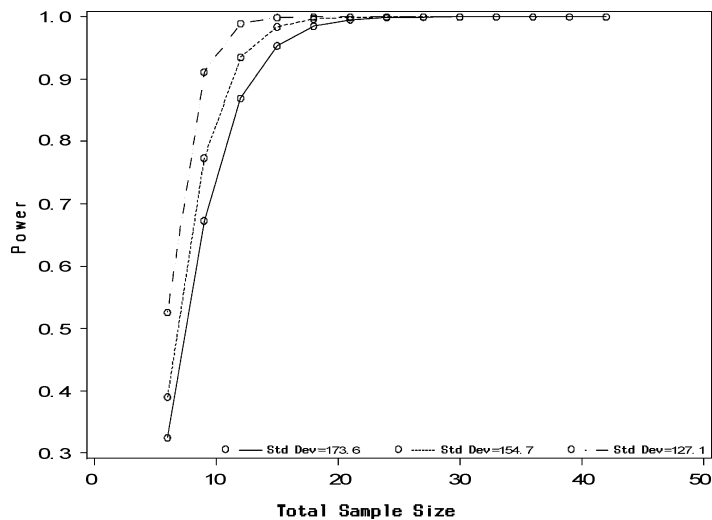


图 6-2 例 6-15 总样本含量与检验功效的关系

分析结果表明：本例的检验效能超过 0.999。

【程序修改指导】 选项“groupmeans = 288.9 | 657.9 | 762.9”定义各水平组指标总体均值的估计值；选项“stddev = 173.6 154.7 127.1”定义三个总体标准差的估计值；选项“groupweights

= (1 1 1)”定义各水平组样本含量的分配比例。语句“plot x = n min = 6 max = 45 key = bycurve (pos = inset numbers = off);”用于画图, 其中的选项“x = n”表示在 x 轴上表示样本含量, 若要在 x 轴上表示检验功效, 则将“x = n”换成“x = power”。上述数字需要根据实际情况修改。

6.3.13 两因素析因设计定量资料方差分析时检验效能的计算

【例 6-16】 某医院用中药复方治疗高胆固醇血症, 把 12 例高胆固醇患者随机分为四组, 用不同疗法治疗。第一组用一般疗法, 第二组在一般疗法上外加用甲药, 第三组在一般疗法上外加用乙药, 第四组在一般疗法上外加用甲药和乙药, 一个月后观察胆固醇降低数 (mg%), 资料如下。问: 若对该资料采用两因素析因设计定量资料的方差分析处理, 检验效能是否足够?

- 第 1 组: 16, 25, 18;
- 第 2 组: 56, 44, 42;
- 第 3 组: 28, 31, 23;
- 第 4 组: 64, 78, 80。

第一步: 先由样本资料估计各试验条件下胆固醇降低数 (mg%) 的总体均值和总体标准差。程序如下, 程序名为: SASTJFX6_16_1.SAS。

```
data step1;
do A=1 to 2;
do B=1 to 2;
do i=1 to 3;
input x @@ ; output;
end;end;end;
cards;
16 25 18
28 31 23
56 44 42
64 78 80
;
run;
proc glm;
class A B;
model x = A B A*B/ss3;
means A*B;
run; quit;
```

主要输出结果:

| Level of | Level of | N | x | |
|----------|----------|---|------------|------------|
| A | B | | Mean | Std Dev |
| 1 | 1 | 3 | 19.6666667 | 4.72581563 |
| 1 | 2 | 3 | 27.3333333 | 4.04145188 |
| 2 | 1 | 3 | 47.3333333 | 7.57187779 |
| 2 | 2 | 3 | 74.0000000 | 8.71779789 |

第二步: 根据上面计算得到的有关估计值进行检验效能分析。程序如下, 程序名为: SAS-TJFX6_16_2.SAS。

```
data step2;
do A=1 to 2;
do B=1 to 2;
input x @@ ;
output;
end;end;
datalines;

proc glmpower data = step2;
class A B;
model x = A|B;
power
stddev = 4.0 8.7
ntotal = 12
power = .;
```

```
19.7 27.3
47.3 74.0
;
run;

plot x = n min = 12 max = 50 key = byfea-
ture (pos = inset);
run;quit;
```

输出结果如下：

| The GLMPower Procedure | | | | |
|--------------------------|--------|---------|---------|--------|
| Fixed Scenario Elements | | | | |
| Dependent Variable | | | x | |
| Total Sample Size | | | 12 | |
| Alpha | | | 0.05 | |
| Error Degrees of Freedom | | | 8 | |
| Computed Power | | | | |
| Index | Source | Std Dev | Test DF | Power |
| 1 | A | 4.0 | 1 | > .999 |
| 2 | A | 8.7 | 1 | > .999 |
| 3 | B | 4.0 | 1 | > .999 |
| 4 | B | 8.7 | 1 | 0.847 |
| 5 | A * B | 4.0 | 1 | 0.950 |
| 6 | A * B | 8.7 | 1 | 0.388 |

例 6-16 总样本含量与检验效能的关系如图 6-3 所示。

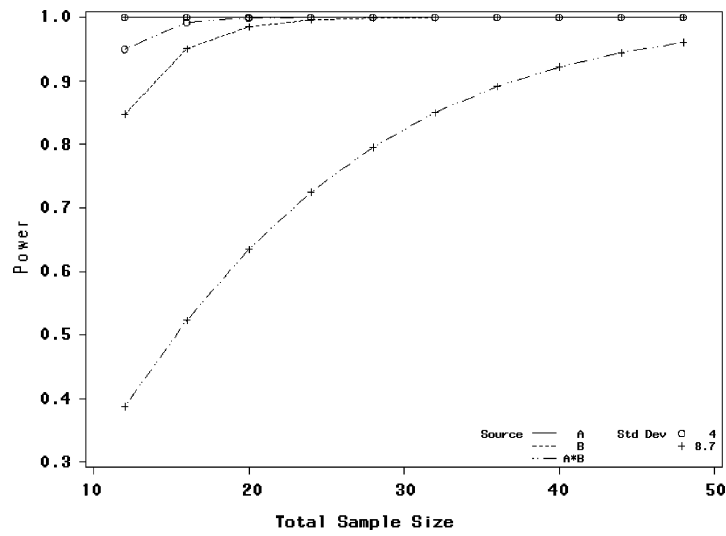


图 6-3 例 6-16 总样本含量与检验效能的关系

分析结果表明：本研究中，要发现甲药和乙药对胆固醇降低数的影响，检验效能可以分别达到 0.999 和 0.847 以上。不过，要发现“甲药用否”与“乙药用否”两个因素之间的交互作用，在总体标准差较大时，检验效能较低：当总体标准差为 8.7 时，检验功效只有 0.388。

【程序修改指导】 对于程序 PGM8_22_H.SAS，data 步用于定义各试验条件下指标的总体均值；glmpower 过程用于功效分析，其中的语句“model x = A | B;”等价于“model x = A B A * B;”，选项“stddev = 4.0 8.7”用于定义指标的总体标准差，可以同时列出多个总体标准差。上述数字需要根据实际情况修改。

6.4 定性资料假设检验中样本含量与检验效能估计

6.4.1 单组设计率的检验时样本含量的估计

【例 6-17】 教育部就《通用规范汉字表》(征求意见稿)公开征求意见,对“琴”、“亲”、“魅”等 44 个汉字的字形拟进行调整,此外,还恢复了 51 个异体字。但新浪网的调查结果(截止到 2009 年 8 月 21 日 9:15,见表 6-2)显示,90.3%的网友表示反对。请问:需询问多少网友才能有 95% 的把握得出持反对意见的比例超过 90% 的结论?取 $\alpha = 0.05$ 。

表 6-2 关于调整 44 个汉字写法的调查结果

| 选 项 | 票 数 | 比例 (%) |
|-----|--------|--------|
| 反对 | 229013 | 90.3 |
| 无所谓 | 12666 | 5.0 |
| 支持 | 11927 | 4.7 |
| 合计 | 253606 | 100.0 |

【分析与解答】 这是一个结果变量为定性变量的单组设计资料的样本含量估计问题。已知检验效能、样本率和总体率的估计值及显著性水平,欲估计达到检验效能的最小样本含量。所需的 SAS 程序如下。程序名为 SASTJFX6_17.SAS。

```
%let p=0.903;%let p0=0.900;
%let delta=0.003;/* delta=p-p0;
*/
%let alpha=0.05;%let beta=0.05;
data d6_17;
n=((probit(1-&alpha))*sqrt(&p0*
(1-&p0))
+(probit(1-&beta))
*sqrt(&p*(1-&p)))/&delta)**2;
n1=int(n)+1;
file print;
PUT #3 @15 '需调查大约' n1 '名网友。';
run;quit;
```

运行结果:需调查大约 106769 名网友。

【程序说明】 5 个 %let 语句分别用于设置样本率、总体率对于类似的问题,只需要根据实际情况对 5 个 %let 语句中的参数进行修改即可。

6.4.2 两样本频率比较时样本含量的估计

【例 6-18】 拟研究两种抗菌药物(其中一种为对照药)对某感染性疾病的治疗效果,经预试验,试验药有效频率为 80%,对照药有效频率为 60%,今要做正式临床试验。问:每组需要观察多少例病人(假设采用双侧检验)?

【分析与解答】 这是一个结果变量为定性变量的成组设计研究的样本含量估计问题。已知两组的样本频率,假定假设检验中容许犯 I、II 型错误的概率已知,试验组与对照组的样本含量之比为 2:1,欲估计最低样本含量。所需的 SAS 程序如下。程序名为 SASTJFX6_18.SAS。

```
Proc power;
TwoSampleFreq test=PChi
GroupProportions=(0.80 0.60)
sides=2
GroupWeights=(2 1)
ntotal=.
Alpha=0.05
Power=0.90;
Run;Quit;
```

运行结果:两组共需 240 例,其中试验组 160 例,对照组 80 例。

【程序说明】 选项“test=PChi”说明处理数据时采用 Pearson χ^2 检验;选项“GroupPropor-

tions = (0.80 0.60)”中的 0.80 和 0.60 分别是两组的有效频率；其他选项的说明同前述各例。程序中的数字需要根据实际情况修改。

6.4.3 多个样本频率比较时样本含量的估计

【例 6-19】 对三种驱肠道蠕虫寄生虫药进行预试验，初估甲药经粪便检查，虫卵阴转频率甲药为 80%，乙药为 85%，丙药为 95%。问：各药经正式临床试验需试验多少病人？

【分析与解答】 这是一个结果变量为二值变量的单因素 3 水平设计试验研究的样本含量估计问题。所需的 SAS 程序如下，程序名为 SASTJFX6_19.SAS。

```
%let alpha=0.05; %let beta=0.10;
%let k=3;
%let p=0.80,0.85,0.95;
%let lam=12.65;
data d6_19;
p1=max(&p);
p2=min(&p);
n=ceil(&lam/(2*(arsin(sqrt(p1))-
arsin(sqrt(0.80)))*2));
file print;
PUT #3 @10 '每组所需样本量为' n '例。';
run;quit;
```

运行结果：每组所需样本量为 112 例。

【程序说明】 第 1 行中， $\alpha=0.05$ 中的 0.05 为可能犯 I 型错误的概率， $\beta=0.10$ 中的 0.10 为可能犯 II 型错误的概率， $k=3$ 中的 3 为组数。第 2 行中， $p=0.80, 0.85, 0.95$ 中的 0.80, 0.85, 0.95 是各组的频率，修改时把各组的频率都写在这里，不同频率之间用逗号分隔， $\text{lam}=12.65$ 中的 12.65 由查 λ 值表得出，上述数字需要根据实际情况修改。

6.4.4 单因素 2 水平设计定性资料等效性检验时检验效能的估计

【例 6-20】 某研究者进行了一项等效性试验，结果如表 6-3 所示。设 $\delta=0.1$ ，取 $\alpha=0.05$ ，假设两组的总体治愈率相同，试分析检验的效能。

【分析与解答】 这是一个成组设计定性资料的等效性检验的检验效能分析问题。 $P = (P_1 + P_2)/2 = (72.41 + 74.71)/2 = 73.56(\%)$ 。所需的 SAS 程序如下，程序名为 SASTJFX6_20.SAS。

```
%let alpha=0.05;
%let delta=0.1;
%let pai_bar=0.7356;
%let n_t=174;
%let k=1;
%let theta=0;
data d6_20;
power = probnorm(sqrt((&theta -
&delta)**2 * &k* &n_t/(&k+1)
/(&pai_bar*(1-&pai_bar))) -
probit(1-&alpha/2)) +
probnorm(sqrt((&theta + &delta)
**2 * &k* &n_t/(&k+1)/(&pai_bar*
(1-&pai_bar)))
-probit(1-&alpha/2)) -1;
file print;
put #3 @15 "Power=" power;
run;quit;
```

运行结果：此等效性检验所能达到的检验效能为 0.123，即 12.3%，远远小于 75%。这一结果表明，该试验的样本含量不足，所得出的结论可信度低。宜加大样本量继续进行试验。

表 6-3 两组药物的疗效

| 药物 | 病例数 | | | 治愈率 (%) |
|-----|-----|-----|-----|------------|
| | 治愈 | 未治愈 | 合计 | |
| 对照药 | 126 | 48 | 174 | 72.41 |
| 试验药 | 130 | 42 | 174 | 74.71 |

【程序说明】 程序中的数字需要根据实际情况修改。对于类似的实际问题，只需修改 6 个“%let”语句中的参数就可以了。

6.4.5 单因素 2 水平设计定性资料非劣效或优效性检验时检验效能的估计

【例 6-21】 对例 6-20，若对资料进行非劣效检验，其他条件保持不变，试分析检验的效能。

【分析与解答】 这是一个成组设计定性资料非劣效性检验的检验效能分析问题。根据已知条件，可用下面的 SAS 程序进行检验效能分析，程序名为 SASTJFX6_21.SAS。

```
%let alpha=0.05;
%let delta_L = -0.1;
%let pai_bar=0.7356;
%let n=174;
%let theta=0;

Datad6_21;
Power=probnorm(abs(&delta_L -
&theta)
* sqrt(&n/(2* &pai_bar* (1 - &pai_
bar)))
-probit(1 - &alpha));
file print;
put #3 @15 "Power=" Power ;
run;quit;
```

运行结果：Power = 0.6808707225。此非劣效检验所能达到的检验效能约为 0.681，即 68.1%，也小于 75%。这一结果表明，该试验的样本含量仍不足，所得出的结论可信度低。宜加大样本量继续进行试验。

【程序说明】 程序中的数字需要根据实际情况修改。对于类似的实际问题，只需修改 5 个“%let”语句中的参数就可以了。

6.4.6 例数相等的两组样本频率比较时检验效能的计算

【例 6-22】 现有某国产抗菌新药需进行临床试验，用同类进口药进行对照，选取患有某种疾病并符合入选条件的病人 60 例，随机分为两组，每组 30 人。试验结果见表 6-4。

两组病人有效率分别为 70.0% 和 90.0%，经 χ^2 检验无统计学意义($P > 0.05$)，此时需考虑到是否由于两组试验病人数不足，致使检验效能偏低，造成假“阴性”结果。试计算其检验效能。

表 6-4 两组病人用药后有效率分析

| 药 物 | 例 数 | 有效率(%) | P |
|-----|-----|--------|-------|
| 试验药 | 30 | 70.0 | >0.05 |
| 对照药 | 30 | 90.0 | |

【分析与解答】 程序名为 SASTJFX6_22.SAS。

```
Proc power;
TwoSampleFreq test = Pchi
GroupProportions =
(0.700 0.900)
nPerGroup = 30

Sides = 2
Alpha = 0.05
Power = .;
Run;Quit;
```

运行结果：检验效能(Power) = 0.490 < 0.75，检验效能不够充分。

【程序说明】 选项“GroupProportions = (0.700 0.900)”中的 0.700 和 0.900 分别是两组的有效率；程序第 4 行指明样本含量，若两组的样本含量不相等，应将选项“nPerGroup = 30”改为“Groups = n1 | n2”的格式；选项“Sides = 2”中的 2 表示双侧检验(为默认状态，可省略)；选

项“Alpha = 0.05”中的 0.05 表示犯 I 型(或假阳性)错误的概率(为默认状态,可省略)。上述数字需要根据实际情况修改。

6.4.7 例数不相等的两组样本频率比较时检验效能的计算

【例 6-23】 某国产新药需对某病种进行临床试验,以考核其疗效,用同类进口药进行对照,试验结果见表 6-5。

表 6-5 两组病人用药后有效率观测结果

| 药 物 | 例 数 | 有效率(%) | P |
|-----|-----|--------|-------|
| 试验药 | 50 | 70.0 | >0.05 |
| 对照药 | 45 | 84.4 | |

【分析与解答】 程序名为 SASTJFX6_23.SAS。

```
Proc power;
TwoSampleFreq test = PChi
GroupProportions = (0.700 0.844)
Groupns = 50|45
Power = .;
Run;Quit;
```

运行结果: 检验效能(Power) = 0.379 < 0.75, 检验效能不够充分。

【程序说明】 参见例 6-22。

6.4.8 单因素 2 水平设计定性资料等效性检验时检验效能的计算

【例 6-24】 某研究者进行了一项等效性试验,结果如表 6-6 所示。设 $\delta = 0.1$, 取 $\alpha = 0.05$, 假设两组的总体治愈率相同, 试分析检验的效能。

表 6-6 两组药物的疗效

| 药物 | 病例数 | | | 治愈率 (%) |
|-----|-----|-----|-----|------------|
| | 治愈 | 未治愈 | 合计 | |
| 对照药 | 126 | 48 | 174 | 72.41 |
| 试验药 | 130 | 42 | 174 | 74.71 |

【分析与解答】 这是一个成组设计定性资料的等效性检验效能分析问题。 $P = (P_1 + P_2)/2 = (72.41 + 74.71)/2 = 73.56(\%)$ 。所需的 SAS 程序如下, 程序名为 SASTJFX6_24.SAS。

```
%let alpha=0.05;
%let delta=0.1;
%let pai_bar=0.7356;
%let n_t=174;
%let k=1;
%let theta=0;
data d6_24;
power=probnorm(sqrt((&theta -
&delta)
** 2* &k* &n_t/(&k+1)/(&pai_bar*
(1 - &pai_bar))) -probit(1 -
&alpha/2)) +
probnorm(sqrt((&theta + &delta)
** 2* &k* &n_t/(&k+1)/(&pai_bar*
(1 - &pai_bar)))
-probit(1 - &alpha/2)) -1;
file print;
put #3 @15 "Power =" power;
run;quit;
```

运行结果: 此等效性检验所能达到的检验效能为 0.123, 即 12.3%, 远远小于 75%。这一结果表明, 该试验的样本含量不足, 所得出的结论可信度低。宜加大样本量继续进行试验。

【程序说明】 程序中的数字需要根据实际情况修改。对于类似的实际问题, 只需修改 6 个“%let”语句中的参数就可以了。

6.4.9 单因素 2 水平设计定性资料非劣效或优效性检验时检验效能的计算

【例 6-25】 对例 6-24，若对资料进行非劣效检验，其他条件保持不变，试分析检验的效能。

【分析与解答】 这是一个成组设计定性资料的非劣效性检验效能分析问题。根据已知条件，可用下面的 SAS 程序进行检验效能分析，程序名为 SASTJFX6_25.SAS。

```
%let alpha=0.05;
%let delta_L=-0.1;
%let pai_bar=0.7356;
%let n=174;
%let theta=0;

Datad6_25;
Power = probnorm (abs (&delta_L -
&theta)
* sqrt (&n/(2* &pai_bar* (1 - &pai_bar)))
-probit(1 - &alpha));
file print;
put #3 @15 "Power =" Power ;
run;quit;
```

运行结果：Power = 0.6808707225。此非劣效检验所能达到的检验效能约为 0.681，即 68.1%，也小于 75%。这一结果表明，该试验的样本含量仍不足，所得出的结论可信度低。宜加大样本量继续进行试验。

【程序说明】 程序中的数字需要根据实际情况修改。对于类似的实际问题，只需修改 5 个“%let”语句中的参数就可以了。

6.5 本章小结

本章介绍了估计样本含量和检验效能的意义、基本概念和必需的前提条件，举例说明如何用 SAS 软件中的编程法和直接用 SAS 中的 POWER 过程及 GLMPOWER 过程实现最常见的 10 余种场合下样本含量及检验效能的估计方法。

(周诗国 陶丽新 刘一松)

第7章 随机化的 SAS 实现

在医学科研实践中，人们经常遇到各种随机化的问题。例如，抽样调查中如何随机选取调查对象；又如，试验设计和临床试验设计中，如何把受试对象随机分配到各试验组中去。实现随机化的方法有很多，简单的随机化可以通过查随机数字表法来实现，复杂一些的随机化常通过 SAS、SPSS 等统计软件来实现。本章着重介绍 SAS 软件中常用的随机抽样和随机分组技巧及原理。

7.1 常见随机抽样和随机分组的种类

随机化应该体现在以下两个方面：(1)随机抽样：每一个符合条件的受试对象参加试验的机会相同，即总体中每个个体有相同的会被抽取进入样本之中；(2)随机分组：每个受试对象被分配到不同组(通常为对照组、不同处理组)中去的机会相同。其中，第(1)条可以保证所得到的样本具有代表性，使试验结论具有普遍意义；第(2)条可以保证各组间受试对象尽可能均衡一致，以提高各组间的可比性。

调查研究中研究者常常采用随机抽样的方法获得样本，这些方法具有统计学理论依据，因此可以对抽样误差进行计算，从而对调查结果的准确性进行评估。常用的随机抽样方法有：单纯随机抽样、分层抽样、整群抽样、系统抽样以及多阶段抽样。

随机分组方法包括完全随机、分层随机、区组随机、分层区组随机、动态随机，一般与盲法合用，有助于避免因处理分配的可预测，在受试者的选择和分组时可能导致的偏倚。实现随机化的方式有多种，可以采用抽签法、查随机数字表或随机排列表、直接由计算机程序来实现。

7.2 调查研究中随机抽样的 SAS 实现

下面介绍几种常用的随机抽样方法。

(1)单纯随机抽样：又称简单随机抽样，它是利用随机的方法从总体中抽取部分抽样单位作为样本，总体中每个抽样单位被抽中的概率相等。单纯随机抽样可以采用掷硬币、掷骰子、抽签或者查随机数字表的方法，也可以在计算机上利用 SAS、SPSS 等软件来实现。单纯随机抽样是最简单的一种随机抽样方法，也是其他抽样方法的基础。

【例 7-1】 表 7-1 给出 20 例病人的基本信息，从中随机抽取 10 名患者作为调查对象。

表 7-1 20 例病人的基本信息

| 病人编号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|------|----|----|----|----|----|----|----|----|----|----|
| 性别 | F | F | M | F | F | F | M | M | M | M |
| 年龄 | 60 | 64 | 37 | 57 | 41 | 31 | 60 | 64 | 58 | 16 |
| 病人编号 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 性别 | M | M | F | F | F | F | F | M | F | F |
| 年龄 | 58 | 63 | 23 | 37 | 20 | 33 | 39 | 40 | 49 | 42 |

用 SAS 实现单纯随机抽样的程序名为 SASTJFX7_1.SAS。

```
data DA7_1;
input id sex $ age;
cards;
1 F 60
2 F 64
.....
20 F 42
;
run;
ods html;
proc surveyselect data=DA7_1 method=srs n=10 out=b;
run;
proc print data=b;
run;
ods html close;
```

【程序说明】 首先，创建 SAS 数据集名为 DA7_1。然后，调用 PROC SURVEYSELECT 进行随机抽样，并指定一些抽样选项，其中“data =”指定输入数据集，用于指定抽样框；“method =”用于指定随机抽样的方法，srs 表示单纯随机抽样；“n =”用于指定抽样大小（此处可以用“rate =”来替换，用于指定抽样率）；“out =”用于指定输出数据集，它包含所有被抽到的样本（可以添加选项“rep = number”来规定重复抽样的次数，如“rep = 2”命令系统同时抽取 2 份样本）。最后用 PROC PRINT 把输出数据集 b 在输出窗口中显示出来。命令 ods html 和 ods html close 用于把程序运行结果以网页格式保存下来。

值得注意的是，此程序采用的是单纯随机抽样（srs），所以，不同时间随机抽样的结果是不同的。

【主要输出结果及解释】

| Obs | id | sex | age |
|-----|----|-----|-----|
| 1 | 3 | M | 37 |
| 2 | 4 | F | 57 |
| 3 | 5 | F | 41 |
| 4 | 6 | F | 31 |
| 5 | 8 | M | 64 |
| 6 | 9 | M | 58 |
| 7 | 11 | M | 58 |
| 8 | 15 | F | 20 |
| 9 | 17 | F | 39 |
| 10 | 19 | F | 49 |

单纯随机抽样的优点是简便易行。缺点是不适合样本量很大的研究，样本量很大时对每个观察对象编号工作量大，有时很难完成；这种方法也不适用于那些个体差异很大的研究对象的抽样，因为在这种情况下，需要在样本量很大时才具有代表性。

(2) 系统抽样：也称等距抽样或机械抽样，它是按照一定的顺序对总体中各抽样单位进行排序，根据样本大小确定抽样间隔，然后随机确定起点，按照抽样间隔抽取每一个单位。它也是在随机抽样的基础上进行的，即要求每次抽样的起点均必须是随机的。该方法操作方便，在对研究总体排序标志有所了解时，可利用某些规律达到较好的抽样效果，故抽样误差通常小于

单纯随机抽样。缺点是，当总体的观察单位按顺序有周期趋势或单调增(减)趋势时，容易产生系统误差，所以该法适用于观察单位分布较均匀的总体。

【例 7-2】 将表 7-1 中的 20 例病人按系统随机抽样的方法抽取 5 名受试者。

用 SAS 实现系统随机抽样的程序名为 SASTJFX7_2.SAS。

```
ods html;
proc surveyselect data=DA7_1 method=sys n=5 out=b;
control id;
run;
proc print data=b;
run;
ods html close;
```

【程序说明】 数据步与例 7-1 相同，此处省略。SURVEYSELECT 过程与例 7-1 的不同之处在于选项 method = sys，规定抽样方法为系统随机抽样法。此外，还多了一个 control 语句，该语句用于指定排序的变量。

值得注意的是，此程序采用的是系统随机抽样(sys)，所以，不同时间随机抽样的结果也是不同的。

【主要输出结果及解释】

| Obs | sex | age | id |
|-----|-----|-----|----|
| 1 | M | 37 | 3 |
| 2 | M | 60 | 7 |
| 3 | M | 58 | 11 |
| 4 | F | 20 | 15 |
| 5 | F | 49 | 19 |

(3) 分层抽样：分层抽样是先按对观测指标影响较大的某种特征，将总体分为若干层，再从每一层内随机抽取一定数量的抽样单位组成样本。分层原则是使层内差异尽可能小，层间的差异尽可能大，从而提高调查的精度。分层抽样按照各层之间的抽样比是否相同，可分为等比例抽样与非等比例抽样。分层抽样的特点是：①调查的精度比较高，因为合理的分层提高了层内同质性，减少了层内的抽样误差；②可以对不同的层采用不同的抽样方法，抽样方法比较灵活；③还可以对各层进行独立分析。

【例 7-3】 对表 7-1 中的病人以性别为分层因素，等比例抽取 10 名受试者。

用 SAS 实现分层随机抽样的程序名为 SASTJFX7_3.SAS。

```
ods html;
proc sort data=DA7_1;
by sex;
run;
proc surveyselect data=DA7_1 method=srs n=(6 4) out=b;
strata sex;
run;
proc print data=b;
run;
ods html close;
```

【程序说明】 数据步与例 7-1 相同，此处省略。过程步，SORT 过程用于将数据集 DA7_1 排序，by 语句指明按性别进行排序，此处排序方式为先 F 后 M。SURVEYSELECT 过程与例 7-1

的不同之处在于选项 $n = (6\ 4)$ ，规定每层中的抽样大小，若每层抽样大小相同，可写为 $n = n_1$ 的形式。此外，还多了一个 STRATA 语句，该语句用于指定分层变量。

值得注意的是，此程序采用的是分层单纯随机抽样 (srs)，所以，不同时间随机抽样的结果也是不同的。

【主要输出结果及解释】

| Obs | sex | id | age | SelectionProb | SamplingWeight |
|-----|-----|----|-----|---------------|----------------|
| 1 | F | 1 | 60 | 0.5 | 2 |
| 2 | F | 2 | 64 | 0.5 | 2 |
| 3 | F | 13 | 23 | 0.5 | 2 |
| 4 | F | 14 | 37 | 0.5 | 2 |
| 5 | F | 19 | 49 | 0.5 | 2 |
| 6 | F | 20 | 42 | 0.5 | 2 |
| 7 | M | 7 | 60 | 0.5 | 2 |
| 8 | M | 8 | 64 | 0.5 | 2 |
| 9 | M | 9 | 58 | 0.5 | 2 |
| 10 | M | 12 | 63 | 0.5 | 2 |

7.3 试验研究中随机分组的 SAS 实现

利用 SAS 软件实现随机分组的途径有很多，目前比较常用的是 SAS/STAT 模块的 PLAN 过程，它可用于构建各种常见的试验设计并对设计方案进行随机化，也可用于产生数字的排列组合表。本节介绍如何利用 PLAN 过程来实现两组和多组的随机分组。

1. 完全随机化分组

完全随机化的一般步骤如下。

- (1) 将受试对象编号，将编号按顺序写成一排。
- (2) 事先规定分组的规则。如分两组时，可规定遇到随机数字为偶数时将对应的受试对象分入试验组、遇到随机数字为奇数时将对应的受试对象分入对照组 (反过来规定也可以)；再如分三组时，可事先规定，凡随机数字除以 3 得余数为 0 者分入第一组，余数为 1 者分入第二组，余数为 2 者分入第三组。当然，也可以规定其他的分组规则，但规则必须事先确定下来，一旦确定不应随意改动。
- (3) 从随机数字表中任意指定的位置开始向后 (或向前) 抄录随机数字，依次写在各编号之下，注意：舍弃不符合要求的随机数字 (如随机数字超过了编号所对应的数字)。
- (4) 根据抄录的随机数字按事先确定的分组规则分组。当各组样本含量不等时最好再用随机的方法进行调整，尽可能使各组的样本含量相等或接近相等 (从统计计算角度看，各组样本含量相等时误差较小)。

【例 7-4】 现有 100 名受试对象，编号为 1 ~ 100 号，试将他们完全随机地均分到试验组和对照组中去。

【分析与解答】 本例属于成组设计，采用完全随机化，对 100 名受试对象进行分组的 SAS 程序为 SASTJFX7_4.SAS。

```
%macro completerandom2 (seed = , num = , p_t = );
proc plan seed = &seed; factors num = &num; output out = a; run;
```



```
data b;set a;no=_n_;if num<=&num* &p_t then group='试验组';
else group='对照组';run;

proc print data=b;var no group;ods html;run;
%mend;

%completerandom2 (seed=20150415,num=100,p_t=0.5);
```

【程序说明】 该程序是一个名为 completerandom2 的宏，其中的宏参数 seed 代表随机种子数；num 代表样本例数；p_t 代表试验组的例数在总例数中所占比例，大部分情况下都是 0.5，或者说大多数研究中两组例数之比是 1:1。对于这段程序的使用者来说，只需要根据自己的实际情况规定上述宏参数的取值就可以了。

【输出结果】 程序 SASTJFX7_4. SAS 的部分分组结果如下：

| Obs | no | group |
|-----|----|-------|
| 1 | 1 | 试验组 |
| 2 | 2 | 试验组 |
| 3 | 3 | 对照组 |
| 4 | 4 | 试验组 |
| 5 | 5 | 试验组 |
| 6 | 6 | 对照组 |
| 7 | 7 | 试验组 |
| 8 | 8 | 试验组 |
| 9 | 9 | 对照组 |
| 10 | 10 | 对照组 |

上述结果中，no 代表受试对象的编号，group 代表受试对象的分组情况。

【例 7-5】 现有 100 名受试对象，编号为 1~100 号，试将他们完全随机地分到试验药物的三个不同剂量组中去，三组的样本例数分别为 50、20、30。

【分析与解答】 本例属于单因素三水平设计，采用完全随机化，对 100 名受试对象进行分组的 SAS 程序为 SASTJFX7_5. SAS。

```
%macro completerandom3 (seed=,num=,p_1=,p_2=);
proc plan seed=&seed;factors num=&num;output out=a;run;

data b;set a;no=_n_;if num<=&num* &p_1 then group='第一组';
else if &num* &p_1<num<=&num* (&p_1+&p_2) then group='第二组';
else group='第三组';run;

proc print data=b;var no group;ods html;run;
%mend;

%completerandom3 (seed=20150415,num=100,p_1=0.5,p_2=0.2);
```

【程序说明】 该程序是一个名为 completerandom3 的宏，其中的宏参数 p_1、p_2 分别代表第一组、第二组的例数在总例数中所占比例，大部分情况下三组例数之比是 1:1:1。其他宏参数的含义可参见前例。对于这段程序的使用者来说，只需要根据自己的实际情况规定上述宏参数的取值就可以了。

【输出结果】 程序 SASTJFX7_5. SAS 的部分分组结果如下：

| Obs | no | group |
|-----|----|-------|
| 1 | 1 | 第一组 |

| | | |
|----|----|-----|
| 2 | 2 | 第一组 |
| 3 | 3 | 第三组 |
| 4 | 4 | 第一组 |
| 5 | 5 | 第一组 |
| 6 | 6 | 第三组 |
| 7 | 7 | 第一组 |
| 8 | 8 | 第一组 |
| 9 | 9 | 第二组 |
| 10 | 10 | 第二组 |

上述结果中，no 代表受试对象的编号，group 代表受试对象的分组情况。

2. 分层随机化分组

分层随机化和区组随机化、分层区组随机化的实现方式是类似的，而且区组随机化、分层区组随机化都可以看成分层随机化的特例，所以本节将以实例介绍区组随机化、分层区组随机化的 SAS 软件实现。读者在进行分层随机化时，仍然使用本节实例中的程序就可以了。

【例 7-6】 试将 240 例某病患者按照区组随机化的方式分配到试验组和对照组，两组患者例数相等。

【分析与解答】 本例属于成组设计，采用区组随机化，对 240 名患者进行分组的 SAS 程序为 SASTJFX7_6.SAS。

```
%macro blockrandom2 (seed=,block=,length=,p_t=);
proc plan seed = &seed; factors block = &block length = &length; output out =
a;run;

data b;set a;no=_n_;if length = <&length* &p_t then group = '试验组';else group
= '对照组';run;

proc print data =b;var no group;ods html;run;
%mend;

%blockrandom2 (seed=20150415,block=40,length=6,p_t=1/2);
```

【程序说明】 该程序是一个名为 blockrandom2 的宏，其中的宏参数 block 代表区组的个数；length 代表区组的长度，一般取值为 4 或 6，区组个数与区组长度之积为样本例数。其他宏参数的含义可参见前例。对于这段程序的使用者来说，只需要根据自己的实际情况规定上述宏参数的取值就可以了。

【输出结果】 程序 SASTJFX7_6.SAS 的部分分组结果如下：

| Obs | no | group |
|-----|----|-------|
| 1 | 1 | 试验组 |
| 2 | 2 | 对照组 |
| 3 | 3 | 试验组 |
| 4 | 4 | 对照组 |
| 5 | 5 | 对照组 |
| 6 | 6 | 试验组 |
| 7 | 7 | 试验组 |
| 8 | 8 | 对照组 |
| 9 | 9 | 对照组 |

| | | |
|----|----|-----|
| 10 | 10 | 试验组 |
| 11 | 11 | 试验组 |
| 12 | 12 | 对照组 |

上述结果中, no 代表受试对象的编号, group 代表受试对象的分组情况。

【例 7-7】 试将 240 例某病患者按照区组随机化的方式分配到试验药物的三个不同剂量组中去, 三组患者例数相等。

【分析与解答】 本例属于单因素三水平设计, 采用区组随机化, 对 240 名患者进行分组的 SAS 程序为 SASTJFX7_7.SAS。

```
%macro blockrandom3(seed=,block=,length=,p_1=,p_2=);
proc plan seed = &seed; factors block = &block length = &length; output out =
a;run;

data b;set a;no=_n_;if length <=&length* &p_1 then group='第一组';
else if &length* &p_1 <length <=&length* (&p_1 +&p_2) then group='第二组';
else group='第三组';run;

proc print data=b;var no group;ods html;run;
%mend;

%blockrandom3(seed=20150415,block=40,length=6,p_1=1/3,p_2=1/3);
```

【程序说明】 该程序是一个名为 blockrandom3 的宏, 其中宏参数的含义可参见前例。对于这段程序的使用者来说, 只需要根据自己的实际情况规定上述宏参数的取值就可以了。

【输出结果】 程序 SASTJFX7_7.SAS 的部分分组结果如下:

| Obs | no | group |
|-----|----|-------|
| 1 | 1 | 第二组 |
| 2 | 2 | 第三组 |
| 3 | 3 | 第一组 |
| 4 | 4 | 第二组 |
| 5 | 5 | 第三组 |
| 6 | 6 | 第一组 |
| 7 | 7 | 第二组 |
| 8 | 8 | 第二组 |
| 9 | 9 | 第三组 |
| 10 | 10 | 第一组 |
| 11 | 11 | 第一组 |
| 12 | 12 | 第三组 |

上述结果中, no 代表受试对象的编号, group 代表受试对象的分组情况。

【例 7-8】 在 4 家医院开展某项临床试验, 试以中心为分层因素, 将 240 例某病患者按照分层区组随机化的方式分配到试验组和对照组。

【分析与解答】 本例属于成组设计, 采用分层区组随机化, 对 240 名患者进行分组的 SAS 程序为 SASTJFX7_8.SAS。

```
%macro strablockrandom2(seed=,center=,block=,length=,p_t=);
proc plan seed = &seed; factors center = &center block = &block length = &length;
output out = a;run;
```

```
data b;set a;no=_n_;if length=<&length* &p_t then group='试验组';else group
='对照组';run;

proc print data=b;var center no group;ods html;run;
%mend;

%strablockrandom2(seed=20150415,center=4,block=10,length=6,p_t=1/2);
```

【程序说明】 该程序是一个名为 strablockrandom2 的宏，其中的宏参数 center 代表分层因素的水平数，临床试验中一般是按中心分层。其他宏参数的含义可参见前例。对于这段程序的使用者来说，只需要根据自己的实际情况规定上述宏参数的取值就可以了。

【输出结果】 程序 SASTJFX7_8.SAS 的部分分组结果如下：

| Obs | center | no | group |
|-----|--------|----|-------|
| 1 | 1 | 1 | 对照组 |
| 2 | 1 | 2 | 试验组 |
| 3 | 1 | 3 | 试验组 |
| 4 | 1 | 4 | 对照组 |
| 5 | 1 | 5 | 试验组 |
| 6 | 1 | 6 | 对照组 |
| 7 | 1 | 7 | 对照组 |
| 8 | 1 | 8 | 对照组 |
| 9 | 1 | 9 | 试验组 |
| 10 | 1 | 10 | 试验组 |
| 11 | 1 | 11 | 对照组 |
| 12 | 1 | 12 | 试验组 |

上述结果中，center 代表中心编号，no 代表受试对象的编号，group 代表受试对象的分组情况。

【例 7-9】 在 4 家医院开展某项临床试验，试以中心为分层因素，将 240 例某病患者按照分层区组随机化的方式分配到试验药物的三个不同剂量组中去，三组患者例数相等。

【分析与解答】 本例属于单因素三水平设计，采用分层区组随机化，对 240 名患者进行分组的 SAS 程序为 SASTJFX7_9.SAS。

```
%macro strablockrandom3(seed=,center=,block=,length=,p_1=,p_2=);
proc plan seed=&seed;factors center=&center block=&block length=&length;
output out=a;run;

data b;set a;no=_n_;if length<=&length* &p_1 then group='第一组';
else if &length* &p_1<length<=&length* (&p_1+&p_2) then group='第二组';
else group='第三组';run;

proc print data=b;var center no group;ods html;run;
%mend;

%strablockrandom3(seed=20150415,center=4,block=10,length=6,p_1=1/3,p_2
=1/3);
```

【程序说明】 该程序是一个名为 strablockrandom3 的宏，其中宏参数的含义可参见前例。对于这段程序的使用者来说，只需要根据自己的实际情况规定上述宏参数的取值就可以了。

【输出结果】 程序 SASTJFX7_9.SAS 的部分分组结果如下：

| Obs | center | no | group |
|-----|--------|----|-------|
| 1 | 1 | 1 | 第三组 |

| | | | |
|----|---|----|-----|
| 2 | 1 | 2 | 第一组 |
| 3 | 1 | 3 | 第二组 |
| 4 | 1 | 4 | 第三组 |
| 5 | 1 | 5 | 第一组 |
| 6 | 1 | 6 | 第二组 |
| 7 | 1 | 7 | 第二组 |
| 8 | 1 | 8 | 第三组 |
| 9 | 1 | 9 | 第一组 |
| 10 | 1 | 10 | 第二组 |
| 11 | 1 | 11 | 第三组 |
| 12 | 1 | 12 | 第一组 |

上述结果中，center 代表中心编号，no 代表受试对象的编号，group 代表受试对象的分组情况。

3. 动态随机分组

动态随机分组的具体做法如下：根据专业知识选取几个拟加以控制的重要非试验因素，假定一个是患者的性别(分为男、女)，另一个是患者的病情(分为轻、中、重)。将先来的两位患者在试验组与对照组各放 1 人，记下他们的性别和病情，记分的方法是每个因素的每个水平出现 1 次记 1 分，计算两组各因素对应水平的得分之差的绝对值，最后求出绝对值之总和，称此“和”为两组患者在两个重要非试验因素上的不平衡指数。若再来第 3 位患者，分别将此患者放入试验组、对照组各 1 次，每次都根据他(或她)的性别、病情的水平取值累加到原有患者的基础之上，可以得到两个不平衡指数，取不平衡指数最小的那种分组方法，这样第 3 位患者的分组就定下来了，用同样的方法去分配以后来的该病患者，直到两组有了事先规定的样本含量时就停止。

【例 7-10】 假定在对肩周炎患者分两组时考虑病情(轻、中、重)、患病时间(短、长)和日常运动量(少、多)三个重要的非试验因素，再假定治疗组和对照组各有了一个患者，他们的基本情况如表 7-2 所示，第 3 位来的是一位新患者，病情重、患病时间短、日常运动量多。问：将此新患者分入哪个组中去为好？

表 7-2 两例患者的基本情况及分组后的平衡程度

| 试验分组 | 例 数 | | | | | | | |
|-------|-----------|---|---|-----------|---|----------|---|----|
| | 病情(轻 中 重) | | | 患病时间(短 长) | | 运动量(少 多) | | 合计 |
| 治疗组 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | |
| 对照组 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | |
| 差的绝对值 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 4 |

注：表中合计值“4”就是“不平衡指数”，下同。

【分析与解答】 试着将新患者分别分入治疗组和对照组，取“不平衡指数”较小者所对应的分配方案，分别参见表 7-3 和表 7-4。

表 7-3 将此名新患者分入治疗组后的平衡程度

| 试验分组 | 例 数 | | | | | | | |
|-------|----------------|---|---|--------------|---|-------------|---|----|
| | 病情(轻 中 重) | | | 患病时间(短 长) | | 运动量(少 多) | | 合计 |
| 治疗组 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | |
| 对照组 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | |
| 差的绝对值 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 5 |

表 7-4 将此名新患者分入对照组后的平衡程度

| 试验分组 | 例 数 | | | | | | | | |
|-------|-----------|---|---|-----------|---|----------|---|----|--|
| | 病情(轻 中 重) | | | 患病时间(短 长) | | 运动量(少 多) | | 合计 | |
| 治疗组 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | | |
| 对照组 | 1 | 0 | 1 | 2 | 0 | 1 | 1 | | |
| 差的绝对值 | 1 | 1 | 1 | 2 | 1 | 0 | 1 | 7 | |

因 $5 < 7$ ，故本例应取表 7-3 作为分配结果，若再有新患者，分配方法相同。
利用 SAS 程序实现上述动态随机化过程，程序名为 SASTJFX7_10.SAS。

```
%let min_home =D:\SASTJFX;
libname min "&min_home";
OPTIONS MSTORED SASMSTORE=min;
* 变量名不能有保留字段 group 和 factor;
%createPatientsList(data=min.aa,vars=condition 3/time 2/exercise 2);
%addPatient(data=min.aa,values=0 1 0/0 1/1 0);
%addPatient(data=min.aa,values=1 0 0/1 0/1 0);
%addPatient(data=min.aa,values=0 0 1/1 0/0 1);
%addPatient(data=min.aa,values=1 0 0/1 0/0 1);
```

【程序说明】 在运行此程序之前，应将一个装有宏程序的文件“sasmacr.sas7bcat”存在 D 盘名为 SASTJFX 的文件夹内。

第一步，根据研究者拟控制的重要非试验因素的个数，修改程序中的第 5 行的表头结构，即修改“vars = ”后面的内容，本例中，第 1 个为病情(condition)，有 3 个水平；第 2 个为患病时间(time)，有 2 个水平；第 3 个为运动量(exercise)，有 2 个水平。若还有其他重要非试验因素，可接着往下写。

第二步，修改或添加数据。每位随机到来的受试者的数据依次写入调用宏 addpatient 的语句中，例如，上述程序中的倒数第 4 句，代表第 1 位受试者在 3 个重要非试验因素上的取值为“0 1 0/0 1/1 0”，其间用“/”隔开；倒数第 3 句，代表第 2 位受试者在 3 个重要非试验因素上的取值为“1 0 0/1 0/1 0”；最后两行代表第 3 位、第 4 位受试者的数据。

第三步，若再有第 5 位、第 6 位，…，可用类似方法将他们的数据写入调用宏 addpatient 语句中去。

第四步，执行程序。先执行前 5 行，后面的调用宏 addpatient 语句可每次执行一行，也可执行多行。

【输出结果】 每次输出结果不仅存在“D:\ SASTJFX”中，其数据集名为 aa，而且，还以网页格式输出到 OUTPUT 窗口。程序中的 4 位受试者动态随机分组结果如下：

| Obs | condition1 | condition2 | condition3 | time1 | time2 | exercise1 | exercise2 | group | factor |
|-----|------------|------------|------------|-------|-------|-----------|-----------|--------|--------|
| 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | group1 | none |
| 2 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | group2 | none |
| 3 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | group1 | 5 VS 7 |
| 4 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | group1 | 6 VS 6 |

上述结果中第 2 列到第 7 列代表患者在非试验因素各个水平上的取值，group 代表患者的分组情况，factor 代表不平衡指数。

7.4 本章小结

本章介绍了常见随机抽样和随机分组的种类, 并对概念及原理进行了简要说明, 举例说明如何用 SAS 软件中的 SURVEYSELECT 过程实现调查研究中的常用随机抽样方法以及使用 PLAN 过程实现试验研究中的多种随机分组方法。

(柳伟伟 胡纯严 毛玮 刘一松)

第3篇 对定量结果进行差异性分析

第8章 单因素设计一元定量资料差异性分析

本章主要介绍利用 SAS 软件中相应的几个过程分析单因素设计一元定量资料的方法，其设计类型包括单组设计、配对设计、成组设计、单因素多水平设计，以及前述四种单因素设计定量资料的实例、用 SAS 实现统计分析结果解释。

8.1 单组设计一元定量资料 t 检验与符号秩和检验

8.1.1 问题与数据

【例 8-1】 已知玉米单交种群 105 的平均穗重为 300g。喷药后，随机抽取 9 个果穗，其穗重分别为 308, 305, 311, 298, 315, 300, 321, 294, 320g。问：喷药后与喷药前的果穗平均重量之间的差别是否具有统计学意义？

【例 8-2】 已知某常规育种水稻平均单株产量 250g，经杂交培育后，随机抽取 10 株，测得单株产量分别为 272, 200, 268, 247, 267, 246, 373, 216, 206, 246g。问：杂交培育的水稻平均单株产量与常规育种水稻平均单株产量之间的差别是否具有统计学意义？

这两个例子基本相同，以例 8-1 来进行详细分析，遇到不同之处再专门说明。

8.1.2 对数据结构的分析

在例 8-1 中，整个资料只涉及一个组，即用某法重复测定果穗重量 9 次，得到 9 个数据，故应属于“单组设计”。指标(果穗重量)为测量得到且可带小数，故该资料属于定量资料，即这是单组设计一元定量资料。

8.1.3 分析目的与统计分析方法的选择

该研究分析目的是考察该 9 个样品所代表的总体均值与标准值之间的差别是否有统计学意义，因此统计分析应该属于单组设计定量资料的统计分析。若定量资料满足独立性和正态分布要求，可进行单组设计一元定量资料 t 检验，此时，可求出该定量指标总体平均值的 95% 置信区间；否则，应进行单组设计一元定量资料符号秩和检验，此时，可基于非参数法求出该定量指标的总体中位数的 95% 置信区间。

8.1.4 SAS 程序中重要内容的说明

例 8-1 资料的程序名为 SASTJFX8_1.SAS。

| 程序 | 说明 |
|---|---|
| DATA SASTJFX8_1; INPUT X @@ ; CARDS; 308 305 311 298 315 300 321 294 320 ; RUN; Ods html; PROC UNIVARIATE NORMAL MU0 =300 /* MU 后为零，不是英文字母 o* / CIBASIC CIPCTLDF alpha =0.05; VARX;RUN; Ods html close; | 建立数据集 以下是输入变量 调用单变量分析过程。该语句中 NORMAL 选项要求对定量资料进行正态性检验，MU0 =300 选项指定标准值为 300，CIBASIC 选项要求基于正态分布分别计算该样本所代表的总体平均值、标准差和方差的 95% 置信区间，CIPCTLDF 要求基于非参数法求置信区间，alpha =0.05 指定置信区间的显著性水准为 0.05，也就是计算 95% 置信区间 |

例 8-2 资料的程序名为 SASTJFX8_2A.SAS，可以与例 8-1 程序完全相同，注意用新数据替换原来的数据，别忘了将 MU0 =300 改为 MU0 =250；但这里还给出另一种形式，供读者选用，设程序名为 SASTJFX8_2B.SAS。

```
DATA SASTJFX8_2B; INPUT X @@ ; Y = (X -250) ; CARDS;  
272 200 268 247 267 246 373 216 206 246  
;  
RUN;  
ODS HTML;  
PROC UNIVARIATE NORMAL; VAR Y; RUN; PROC MEANS; VAR X; RUN;  
ODS HTML CLOSE;
```

注意：“Y = X -250;”是一个赋值语句，应根据具体情况修改标准值 250。

8.1.5 主要分析结果及解释

| 位置检验：Mu0 = 300 | | | | |
|----------------|-----------|-----------|-----------|--------|
| 检验 | 统计量 | P 值 | | |
| 学生 t | t | 2.495401 | Pr > t | 0.0372 |
| 符号 | M | 2 | Pr >= M | 0.2891 |
| 符号秩 | S | 14 | Pr >= S | 0.0547 |
| 正态性检验 | | | | |
| 检验 | 统计量 | P 值 | | |
| Shapiro-Wilk | W | 0.954097 | Pr < W | 0.7350 |
| 基本置信限正态假设 | | | | |
| 参数 | 估计值 | 95% 置信限 | | |
| 均值 | 308.00000 | 300.60719 | 315.39281 | |
| 标准偏差 | 9.61769 | 6.49634 | 18.42529 | |
| 方差 | 92.50000 | 42.20240 | 339.49147 | |

对例 8-1 而言，首先看关于正态性检验的结果。W =0.954097、P =0.7350 >0.05，可以认为这组定量数据服从正态分布；再看 t 检验的结果，并结合统计学知识和专业知识给出统计学结论

和专业结论。本例， $t = 2.495401$ 、 $P = 0.0372$ ，故按 $\alpha = 0.05$ 水准，拒绝 $H_0(\mu_0 = 300)$ ，接受 $H_1(\mu_0 \neq 300)$ ，可以认为喷药后果穗重量的均值与标准值(300g)之间的差别有统计学意义。

专业结论：因 $\bar{x} = 308\text{g}$ ，标准值为 300g，结合统计学结论，可以认为喷药后果穗重量的均值高于标准值。求得总体平均值的 95% 置信区间为(300.6, 315.4)g。

对例 8-2 而言，由 SASTJFX8_2B. SAS 程序输出的主要结果如下：

| 位置检验：Mu0 = 0 | | | | |
|--------------|-----|----------|----------|--------|
| 检验 | 统计量 | P 值 | | |
| 学生 t | t | 0.264099 | Pr > t | 0.7977 |
| 符号 | M | -1 | Pr >= M | 0.7539 |
| 符号秩 | S | -2.5 | Pr >= S | 0.8262 |
| 正态性检验 | | | | |
| 检验 | 统计量 | P 值 | | |
| Shapiro-Wilk | W | 0.836851 | Pr < W | 0.0404 |

首先查看关于变量 Y 的正态性检验的结果，有 $W = 0.836851$ 、 $P = 0.0404 < 0.05$ ，故可以认为这组定量资料不服从正态分布；接着就应查看符号秩和检验的结果，即位置检验：Mu0 = 0 的最后一行，并结合统计学知识和专业知识给出统计学结论和专业结论。本例， $S = -2.5$ 、 $P = 0.8262 > 0.05$ ，故按 $\alpha = 0.05$ 水准，认为此样本所代表的总体中位数与给定的标准值 250g 之间的差别无统计学意义。

专业结论：可认为该类水稻杂交培育后的单株产量的总体中位数与常规育种的平均单株产量(250g)接近相等。求得中位数为 246.5g，其 95% 置信区间为(206.0, 272.0)g。注意，关于中位数及其置信区间需查看本例第 1 个程序输出的结果中“分位数”部分。本例中由于变量 X 并不服从正态分布，所以描述其平均水平时使用中位数，相应的需要计算中位数的 95% 可信区间，由于输出结果中关于分位数置信区间部分的内容较多，此处从略。

8.2 配对设计一元定量资料 t 检验与符号秩和检验

8.2.1 问题与数据

【例 8-3】 对血小板活化模型大鼠以 ASA 进行实验性治疗，以血浆 TXB₂(ng/L)为指标，其结果见表 8-1。试进行统计分析。

【例 8-4】 一位社会学者为了判定聪明人选择的配偶是否也聪明，随机选取 32 对夫妻，并测得智商得分(IQ)，见表 8-2。问：配偶之间差别是否有统计学意义？

表 8-1 大鼠血小板活化模型 ASA 治疗前后血浆 TXB₂ 的变化 (ng/L)

| 大鼠号 | 血浆 TXB ₂ (ng/L) | |
|-----|----------------------------|-----|
| | 给药前 | 给药后 |
| 1 | 250 | 184 |
| 2 | 226 | 205 |
| ⋮ | ⋮ | ⋮ |
| 10 | 176 | 176 |

表 8-2 32 对夫妻 IQ 得分

| 配偶号 | IQ 得分 | |
|-----|-------|-----|
| | 丈夫 | 妻子 |
| 1 | 95 | 95 |
| 2 | 103 | 98 |
| ⋮ | ⋮ | ⋮ |
| 32 | 96 | 102 |

8.2.2 对数据结构的分析

在例 8-3、例 8-4 中整个资料涉及一个试验因素的 2 个水平，且在这 2 个水平作用下获得的相同指标是成对出现的，每一对中的 2 个数据来自于同一个个体或条件相近的两个个体。故应属于“配对设计”。

8.2.3 分析目的与方法选择

对于例 8-3，研究分析目的是考察该 10 对大鼠给药前后血浆 TXB₂ 差值所代表总体差值与 0 之间的差别是否有统计学意义，因此统计分析应该属于配对设计定量资料的统计分析。若定量资料满足独立性和正态分布要求，可进行单组设计一元定量资料 t 检验，此时，可求出该定量指标的总体平均值的 95% 置信区间；否则，应进行配对设计一元定量资料符号秩和检验，此时，可基于非参数法求出该定量指标的总体中位数的 95% 置信区间。例 8-4 数据经正态性检验，不符合正态性，因此选用符号秩和检验。

8.2.4 SAS 程序中重要内容的说明

例 8-3 资料的程序名为 SASTJFX8_3.SAS，此处仅对例 8-3 数据进行分析。例 8-4 定量资料的程序名为 SASTJFX8_4.SAS，详细程序与例 8-3 基本一致。

| 程序 | 说明 |
|---|--|
| DATA SASTJFX8_3; INPUT x1 x2 @@ ; d = x2 - x1; CARDS; 250 184 226 205 180 182 356 248 280 196 210 204 276 214 326 274 208 200 176 176 ; RUN; Ods html; PROC UNIVARIATE NORMAL CIBASIC CIPCTLDF alpha = 0.05; VAR d; RUN; Ods html close; | 建立数据集 以下是输入变量 d 为 x2、x1 的差值 对差值 d 进行单变量分析，并对 d 进行正态性检验。该语句中 NORMAL 选项要求对定量资料进行正态性检验，CIBASIC 选项要求分别计算该样本所代表的总体平均值、标准差和方差的 95% 置信区间，CIPCTLDF 要求基于非参数法求置信区间，alpha = 0.05 指定置信区间的显著性水准为 0.05，也就是计算 95% 置信区间 |

8.2.5 主要分析结果及解释

| | | | | |
|---------------|-----|----------|----------|--------|
| UNIVARIATE 过程 | | | | |
| 位置检验: Mu0 = 0 | | | | |
| 位置检验: Mu0 = 0 | | | | |
| 检验 | 统计量 | P 值 | | |
| 学生 t | t | -3.27465 | Pr > t | 0.0096 |
| 符号 | M | -3.5 | Pr >= M | 0.0391 |
| 符号秩 | S | -21.5 | Pr >= S | 0.0078 |

正态性检验

| 检验 | 统计量 | P 值 |
|--------------|-----|----------|
| Shapiro-Wilk | W | 0.902699 |
| | | Pr < W |
| | | 0.2345 |

对于例 8-3, 首先查验正态性检验的结果。d 变量正态性检验的结果: $W = 0.902699$ 、 $P = 0.2345 > 0.05$, 可看出差值符合正态分布, 故选用 t 检验的结果, $t = -3.27465$, $P = 0.0096$, 故按 $\alpha = 0.05$ 水准, 认为平均来说给药前后血浆 TXB_2 (ng/L) 之间的差别有统计学意义。

专业结论: 因给药后与给药前的血浆 TXB_2 差值的平均值 -40.5000 小于 0, 结合统计学结论, 可认为 ASA 药物可降低大鼠血浆 TXB_2 (ng/L) 的含量。

位置检验: $\mu_0 = 0$

| 检验 | 统计量 | P 值 |
|------|-----|----------|
| 学生 t | t | 0.630192 |
| | | Pr > t |
| | | 0.5332 |
| 符号 | M | 6 |
| | | Pr >= M |
| | | 0.0428 |
| 符号秩 | S | 85.5 |
| | | Pr >= S |
| | | 0.0778 |

正态性检验

| 检验 | 统计量 | P 值 |
|--------------|-----|----------|
| Shapiro-Wilk | W | 0.820765 |
| | | Pr < W |
| | | 0.0001 |

对于例 8-4, 首先查验正态性检验的结果, d 变量正态性检验的结果: $W = 0.820765$, $P = 0.0001 < 0.05$, 可看出差值不符合正态分布, 故选用符号秩和检验的结果, 本例, $S = 85.5$, $P = 0.0778 > 0.05$, 故按 $\alpha = 0.05$ 水准, 认为平均来说丈夫与妻子 IQ 得分接近相等。

8.3 成组设计一元定量资料 t 检验

8.3.1 问题与数据

【例 8-5】 一个小麦新品种经过 6 代选育, 从第 5 代(A 组)中抽出 10 株, 株高为 66、65、66、68、62、65、63、66、68、62 (cm), 又从第 6 代(B 组)中抽出 10 株, 株高为 64、61、57、65、65、63、62、63、64、60 (cm)。问: 株高性状是否已经达到稳定?

8.3.2 对数据结构的分析

在例 8-5 中, 整个资料涉及 2 个组, 每组随机抽取 10 个数据, 测量指标为“株高”, 故应属于“成组设计”, 即资料类型为成组设计一元定量资料。

8.3.3 分析目的与方法选择

该研究分析目的是考察 2 组之间的总体均值差别是否有统计学意义, 因此统计分析应该属于成组设计定量资料的统计分析。若定量资料满足独立性、正态分布和方差齐性的要求, 可进行成组设计一元定量资料 t 检验, 此时, 可求出该定量指标的总体平均值的 95% 置信区间; 否则, 应进行单组设计一元定量资料符号秩和检验。

8.3.4 SAS 程序中重要内容的说明

例 8-5 资料的程序名为 SASTJFX8_5.SAS。

| 程序 | 说明 |
|-------------------------------|-------------------------------|
| DATA SASTJFX8_5; | 建立数据集 |
| INPUT g \$ n; DO i=1 to n; | |
| INPUT x @@ ; | 以下是输入数据 |
| OUTPUT; END; | |
| CARDS; | |
| A 10 | |
| 66 65 66 68 62 65 63 66 68 62 | |
| B 10 | |
| 64 61 57 65 65 63 62 63 64 60 | |
| ; | |
| RUN; | |
| Ods html; | 以 g 作为分组变量排序 |
| PROC SORT; | 调用单变量过程, 进行正态性检验。选项 NOR- |
| BY g; | MAL 表示进行正态性检验。 |
| RUN; | 进行 t 检验。选项 COCHRAN 表示输出 COCH- |
| PROC UNIVARIATE NORMAL; | RAN 近似 t 检验的结果 |
| VAR x; | |
| BY g; | |
| RUN; | |
| PROC TTEST COCHRAN; | |
| CLASS g; | |
| VAR x; | |
| RUN; | |
| Ods html close; | |

8.3.5 主要分析结果及解释

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---------------------|------|----------|------------------|-----------|------------------|---------------------|---------|---------------------|---------|---------|---------|--|--|--|--|--|--|--|--|--|--|--|--|--|
| g = A | | | | | | | | | | | | | | | | | | | | | | | | |
| 正态性检验 | | | | | | | | | | | | | | | | | | | | | | | | |
| 检验 | | 统计量 | | P 值 | | | | | | | | | | | | | | | | | | | | |
| Shapiro-Wilk | W | 0.902418 | Pr < W | 0.2329 | | | | | | | | | | | | | | | | | | | | |
| g = B | | | | | | | | | | | | | | | | | | | | | | | | |
| 正态性检验 | | | | | | | | | | | | | | | | | | | | | | | | |
| 检验 | | 统计量 | | P 值 | | | | | | | | | | | | | | | | | | | | |
| Shapiro- Wilk | W | 0.899644 | Pr < W | 0.2171 | | | | | | | | | | | | | | | | | | | | |
| The TTEST Procedure | | | | | | | | | | | | | | | | | | | | | | | | |
| Statistics | | | | | | | | | | | | | | | | | | | | | | | | |
| Variable | g | N | Lower CL Mean | Mean | Upper CL Mean | Lower CL Std Dev | Std Dev | Upper CL Std Dev | Std Err | Minimum | Maximum | | | | | | | | | | | | | |
| x | A | 10 | 63.538 | 65.1 | 66.662 | 1.5017 | 2.1833 | 3.9858 | 0.6904 | 62 | 68 | | | | | | | | | | | | | |
| x | B | 10 | 60.609 | 62.4 | 64.191 | 1.7219 | 2.5033 | 4.5701 | 0.7916 | 57 | 65 | | | | | | | | | | | | | |
| x | Diff | | 0.4932 | 2.7 | 4.9068 | 1.7748 | 2.3488 | 3.4734 | 1.0504 | | | | | | | | | | | | | | | |
| (1 - 2) | | | | | | | | | | | | | | | | | | | | | | | | |
| T-Tests | | | | | | | | | | | | | | | | | | | | | | | | |
| Variable | | Method | | Variances | | DF | t Value | Pr > t | | | | | | | | | | | | | | | | |
| x | | Pooled | | Equal | | 18 | 2.57 | 0.0193 | | | | | | | | | | | | | | | | |

| | | | | | |
|-----------------------|---------------|---------|--------|---------|--------|
| x | Satterthwaite | Unequal | 17.7 | 2.57 | 0.0194 |
| x | Cochran | Unequal | 9 | 2.57 | 0.0302 |
| Equality of Variances | | | | | |
| Variable | Method | Num DF | Den DF | F Value | Pr > F |
| x | Folded F | 9 | 9 | 1.31 | 0.6902 |

首先查验正态性检验的结果，可知两组均符合正态分布（W 分别为 0.902418、0.899644；P 分别为 0.2329、0.2171，均大于 0.05），而后查验两组方差齐性检验的结果，可知两组总体方差相等（F = 1.31、P = 0.6902 > 0.05）。故采用成组设计定量资料的 t 检验：t = 2.57、P = 0.0193 < 0.05，两个平均值之间的差别有统计学意义。

专业结论：此小麦品种第 5 代平均株高高于第 6 代平均株高，株高性状没有达到稳定。

8.4 成组设计一元定量资料两种近似 t 检验和 Wilcoxon 秩和检验

8.4.1 问题与数据

【例 8-6】探讨正己烷职业接触人群生化指标特征，用气相色谱法检测受检者尿液 2，5 - 己二酮浓度（mg/L），为该人群的健康监护寻找动态观察依据。正己烷职业接触组（A 组）为广州市印刷行业彩印操作位作业人员 64 人，其均在同一个大的车间轮班工作，工作强度相当；对照组（B 组）选同厂其他车间工人 53 人。两组人员除接触正己烷因素不同外，生活水平、生活习惯、劳动强度、吸烟、饮酒情况基本相同。问：两组间尿液中 2，5 - 己二酮浓度（mg/L）平均含量之间的差别是否有统计学意义？数据如下所示。

正己烷职业接触组：

2.89、1.85、2.27、2.07、1.62、1.77、2.53、2.02、2.07、2.07、1.93、3.01、1.93、1.88、1.55、1.36、2.23、2.55、1.73、2.65、1.95、2.45、1.41、2.46、2.38、1.55、2.16、2.01、1.37、2.16、2.00、2.07、2.57、2.11、2.37、1.39、2.18、2.33、1.46、2.16、2.03、2.96、2.21、2.00、2.58、2.19、2.41、1.68、1.93、1.93、1.93、1.87、1.74、2.70、1.83、2.17、2.52、2.09、2.28、1.65、1.19、1.58、0.89、1.65

对照组：

0.27、0.36、0.26、0.16、0.49、0.58、0.16、0.45、0.22、0.25、0.66、0.05、0.31、0.12、0.51、0.30、0.37、0.14、0.28、0.33、0.36、0.51、0.37、0.36、0.47、0.34、0.72、0.39、0.55、0.17、0.27、0.33、0.30、0.26、0.50、0.17、0.22、0.18、0.17、0.62、0.27、0.26、0.34、0.17、0.61、0.42、0.39、0.28、0.36、0.43、0.24、0.15、0.19

8.4.2 对数据结构的分析

在例 8-6 中，整个资料涉及 2 个组，观测指标为“尿液 2，5 - 己二酮浓度（mg/L）”，故应属于“成组设计（或单因素 2 水平设计）”，即资料类型为成组设计一元定量资料。

8.4.3 分析目的与统计分析方法的选择

该研究分析目的是考察 2 组之间的总体均值差别是否有统计学意义，因此统计分析应该属于成组设计定量资料的统计分析。若定量资料满足独立性、正态分布和方差齐性的要求，可

进行成组设计一元定量资料 t 检验，此时，可求出该定量指标的总体平均值的 95% 置信区间；否则，应进行单组设计一元定量资料符号秩和检验。本例经过检验不符合参数检验的前提条件，故应选择非参数检验法，即秩和检验。

8.4.4 SAS 程序中重要内容的说明

例 8-6 资料的程序名为 SASTJFX8_6.SAS。

| 程序 | 说明 |
|--|---|
| DATA SASTJFX8_6; INPUT g \$ n;DO i=1 to n; INPUT x@@ ; OUTPUT; END; CARDS; A64 2.89 1.85...1.65 B 53 0.27 0.36...0.19 ; RUN; ods html; PROC SORT; BY g; RUN; PROC UNIVARIATE NORMAL; VAR x; BY g;RUN; PROC TTEST COCHRAN; CLASS g; VAR x;RUN; PROC NPAR1WAY WILCOXON; CLASS g; VAR x; RUN; ods html close; | 建立数据集 以下是输入变量 以下是输入数据 (这里为节省篇幅，将很多数据省略了，实际运用 SAS 时，数据必须完整，下同，不再赘述) 以 g 作为分组变量排序 调用单变量过程，进行正态性检验，选项 NORMAL 表示进行正态性检验 调用 ttest 过程进行 t 检验、近似 t 检验，选项 COCHRAN 表示输出 COCHRAN 近似 t 检验的结果。进行秩和检验。选项 WILCOXON 表示只输出 Wilcoxon 秩和检验的结果 |

8.4.5 主要分析结果及解释

| | | | | |
|---------------|-----|----------|--------|--------|
| g = A | | | | |
| 正态性检验 | | | | |
| 检验 | 统计量 | P 值 | | |
| Shapiro- Wilk | W | 0.992748 | Pr < W | 0.9714 |
| g = B | | | | |
| 正态性检验 | | | | |
| 检验 | 统计量 | P 值 | | |
| Shapiro- Wilk | W | 0.963328 | Pr < W | 0.1030 |

以上结果表明：两组定量资料均满足正态分布要求。
说明：此处有关于两组的基本统计量(包括均值及其置信区间)的计算结果，因表格过宽，从略。

| T-Tests | | | | | |
|----------|---------------|-----------|------|---------|---------|
| Variable | Method | Variances | DF | t Value | Pr > t |
| x | Pooled | Equal | 115 | 27.58 | < .0001 |
| x | Satterthwaite | Unequal | 81.1 | 29.70 | < .0001 |
| x | Cochran | Unequal | . | 29.70 | 0.0001 |

上面的第一行为对定量资料进行成组设计一元定量资料 t 检验的结果，仅当两组定量资料满足独立性、正态性和方差齐性要求时，才能用此行结果对两均值之间差别是否具有统计学意义下结论。

上面的第二行和第三行为对定量资料进行成组设计一元定量资料两种近似 t 检验的结果（分别为对自由度进行校正的 Satterthwaite 法和对临界值进行校正的 Cochran 法，相对来说，前一种校正法准确度高一些），仅当两组定量资料满足独立性、正态性但不满足方差齐性要求时，才能用此两行之一的结果对两均值之间差别是否具有统计学意义下结论。

| Equality of Variances | | | | | |
|-----------------------|----------|--------|--------|---------|---------|
| Variable | Method | Num DF | Den DF | F Value | Pr > F |
| x | Folded F | 63 | 52 | 8.07 | < .0001 |

以上是对本例定量资料进行方差齐性检验所得到的结果。结果表明：两组定量资料不满足方差齐性要求。

| Wilcoxon Scores(Rank Sums) for Variable x | | | | | |
|---|----|---------------|-------------------|------------------|------------|
| Classified by Variable g | | | | | |
| g | N | Sum of Scores | Expected Under H0 | Std Dev Under H0 | Mean Score |
| A | 64 | 5472.0 | 3776.0 | 182.607282 | 85.50 |
| B | 53 | 1431.0 | 3127.0 | 182.607282 | 27.00 |
| Average scores were used for ties. | | | | | |

以上是秩和检验之前给出的基本统计量的计算结果，其中，最后一列平均得分就是平均秩，便于下专业结论。

| Wilcoxon Two-Sample Test | |
|--|-----------|
| Statistic | 1431.0000 |
| Normal Approximation | |
| Z | -9.2850 |
| One-Sided Pr < Z | < .0001 |
| Two-Sided Pr > Z | < .0001 |
| t Approximation | |
| One-Sided Pr < Z | < .0001 |
| Two-Sided Pr > Z | < .0001 |
| Z includes a continuity correction of 0.5. | |

| Kruskal-Wallis Test | |
|---------------------|---------|
| Chi-Square | 86.2612 |
| DF | 1 |
| Pr > Chi-Square | < .0001 |

首先查验正态性检验的结果，可知两组均符合正态分布，（W 值分别为 0.992748、0.963328；P 值分别为 0.9714、0.1030，均大于 0.05）。两组方差齐性检验的结果是方差不齐

($F=8.07$, $P<0.0001$)。故可以直接选用非参数检验的结果: 本例采用成组设计定量资料的秩和检验: $Z=-9.2850$, $P=0.0001$, 两组尿液中 2, 5-己二酮浓度平均含量之间的差别有统计学意义。当然, 也可以选用两种近似 t 检验的结果之一来下结论, 与秩和检验法所得结论相同, 从略。最后给出的是 χ^2 近似计算的结果(注: 在 SAS 9.1 之前, 仅在单因素多水平设计一元定量资料秩和检验时才给出 χ^2 近似计算的结果)。

专业结论: 正己烷职业接触组尿液中 2, 5-己二酮平均浓度为 2.04mg/L(平均秩 85.50), 对照组尿液中 2, 5-己二酮平均浓度为 0.33 mg/L(平均秩 27.00), 结合统计学结论可认为正己烷职业接触组尿液中 2, 5-己二酮平均浓度高于对照组尿液中 2, 5-己二酮平均浓度。

8.5 成组设计一元定量资料三种特殊的比较 ——优效性、非劣效性和等效性 t 检验

8.5.1 何为三种特殊的假设检验

对于单因素 2 水平设计一元定量资料, 人们常用成组设计一元定量资料的 t 检验或者 Wilcoxon 秩和检验进行假设检验。事实上, 这里提及的常用的假设检验属于一般的差异性检验方法。然而, 在新药和医疗器械临床试验研究中, 还有三种特殊的假设检验方法, 它们分别被称为优效性检验、非劣效性检验和等效性检验。

优效性检验指主要研究目的是显示试验药的治疗效果优于对照药(安慰剂对照或阳性对照)的试验。在试验设计阶段需要设定一个界值 δ_U , 来界定试验药的优效性。

非劣效性检验指主要研究目的是显示试验药的治疗效果在临床上不比阳性对照药差的试验。在试验设计阶段需要设定一个界值 δ_L , 来界定试验药是否不比阳性对照药疗效差过预先设定的这个界值。

等效性检验指主要研究目的是显示两种治疗效果之间的差别大小在临床上并无重要意义的试验。在试验设计阶段需要设定等效性界值(δ_L, δ_U), 来界定两种治疗的等效性。

8.5.2 成组设计一元定量资料优效性检验

1. 与优效性检验有关的问题与数据

【例 8-7】 评价消痰降脂胶囊治疗高脂血症的有效性。选用高脂血症患者 82 例, 采用随机分组、双盲、多中心、安慰剂组平行对照设计, 进行了消痰降脂胶囊(5 粒/次)与安慰剂组的疗效比较, 用药 8 周测量胆固醇的降低值。如果消痰降脂胶囊组的胆固醇降低值超过安慰剂组 0.28mmol/L, 才能认为消痰降脂胶囊优效于安慰剂。资料见表 8-3, 试评价消痰降脂胶囊是否优效于安慰剂。

表 8-3 消痰降脂胶囊组和安慰剂组的胆固醇降低值 (mmol/L)

| 药物种类 | n | \bar{x} | s |
|--------|----|-----------|------|
| 消痰降脂胶囊 | 39 | 1.25 | 1.13 |
| 安慰剂 | 43 | 0.64 | 0.51 |

2. 如何用 SAS 实现优效性检验

【分析与解答】 本临床试验的目的是评价消痰降脂胶囊降低胆固醇的效果是否优于安慰剂, 并且设定了优效性的界值, 这时应采用优效性检验。应用 SAS 进行分析, 程序如下, 程序名为 SASTJFX8_7.SAS。

```
data SASTJFX8_7;
  n1 = 39; n2 = 43;          /* 第1步 */
  mean1 = 1.25; mean2 = 0.64; /* 第2步 */
  s1 = 1.13; s2 = 0.51;      /* 第3步 */
  U = 0.28;                  /* 第4步 */
  ss1 = s1**2 * (n1 - 1);    /* 第5步 */
  ss2 = s2**2 * (n2 - 1);    /* 第6步 */
  sc = (ss1 + ss2) / (n1 + n2 - 2); /* 第7步 */
  se = sqrt(sc * (1/n1 + 1/n2)); /* 第8步 */

  t = ((mean1 - mean2) - U) / se; /* 第9步 */
  p = 1 - probt(t, n1 + n2 - 2); /* 第10步 */

run;
ods html;
proc print;                  /* 第11步 */
  var t p;
run;
ods html close;
```

【程序说明】 data 步建立数据集 SASTJFX8_7，第1步定义消痰降脂胶囊组和安慰剂组的样本含量；第2步定义消痰降脂胶囊组和安慰剂组的样本平均数；第3步定义消痰降脂胶囊组和安慰剂组的样本标准差；第4步定义优效性检验的界值 U；第5步和第6步分别定义消痰降脂胶囊组和安慰剂组的样本平均数的离均差平方和；第7步计算两组的合并方差；第8步计算两样本平均数之差的标准误；第9步对 H_0 ，即优效性界值 U 进行假设检验，计算检验统计量 t；第10步计算 t 值对应的 t 分布右侧的累计概率；第11步输出 t 值和 p 值。

【主要输出结果及解释】

| Obs | t | p |
|-----|---------|----------|
| 1 | 1.73123 | 0.043633 |

【统计与专业结论】 $t = 1.73123$ ， $p = 0.043633$ ，按照 $\alpha = 0.05$ ，拒绝 H_0 ，接受 H_1 ，可以认为消痰降脂胶囊治疗高脂血症的疗效优于安慰剂。

8.5.3 成组设计一元定量资料非劣效性检验

1. 与非劣效性检验有关的问题与数据

【例 8-8】 观察金童颗粒治疗小儿抽动障碍肾阴亏损、肝风内动证的有效性，采用随机、阳性药平行对照、双盲双模拟的方法。选取 452 例受试对象，按 3:1 分为试验组和对照组，两组分别用金童颗粒和泰必利片，疗程 6 周。资料如表 8-4 所示。如果金童颗粒组的运动性抽动积分下降值低于阳性对照组 1.77 分，就认为该药没有推广价值。试分析金童颗粒对小儿抽动障碍的治疗效果是否不劣于泰必利。

表 8-4 试验组和对照组的运动性抽动积分下降值

| 药物种类 | n | \bar{x} | s |
|------|-----|-----------|------|
| 金童颗粒 | 339 | 8.58 | 3.97 |
| 泰必利 | 113 | 9.12 | 3.61 |

2. 如何用 SAS 实现非劣效性检验

【分析与解答】 该资料属于成组设计一元定量资料，本临床试验的目的是分析金童颗粒对小儿抽动障碍的治疗效果是否不劣于泰必利，设定非劣效性界值为 -1.77，使用非劣效性检验分析该资料。应用 SAS 进行分析，程序如下，程序名为 SASTJFX8_8.SAS。

```
data SASTJFX8_8;
  n1 = 339; n2 = 113;
  mean1 = 8.58; mean2 = 9.12;
  s1 = 3.97; s2 = 3.61;
  L = -1.77;          /* 第1步 */

  t = ((mean1 - mean2) - L) / se; /* 第2步 */
  p = 1 - probt(t, n1 + n2 - 2);

run;
ods html;
```

```

ss1 = s1 ** 2 * (n1 - 1); ss2 = s2 ** 2 *
* (n2 - 1);
sc = (ss1 + ss2) / (n1 + n2 - 2);
se = sqrt(sc * (1/n1 + 1/n2));

proc print;
var t p;
run;
ods html close;

```

【程序说明】 第1步定义非劣效性界值 L ；第2步对 H_0 ，即非劣效性界值 L 进行假设检验，计算 t 值。

【主要输出结果及解释】

| Obs | t | p |
|-----|---------|------------|
| 1 | 2.91574 | .001862826 |

【统计与专业结论】 $t = 2.91574$, $p = 0.001862826$, 按照 $\alpha = 0.05$, 拒绝 H_0 , 接受 H_1 , 认为金童颗粒对小儿抽动障碍的治疗效果不劣于泰必利, 具有推广价值。

8.5.4 成组设计一元定量资料等效性检验

1. 与等效性检验有关的问题与数据

【例8-9】 某研究者观察氯沙坦与伊贝沙坦治疗对伴高尿酸血症的原发性高血压患者血清尿酸水平的影响并评价其降压疗效。采用多中心、随机、双盲、平行对照设计。随机选取320例受试者, 治疗6周后收缩压改变值的情况见表8-5, 根据临床经验, 设定等效性界值为5mmHg, 试评价两种药物的降压效果是否等效。

表8-5 两组患者治疗6周后收缩压下降幅度 (mmHg)

| 药物种类 | n | \bar{x} | s |
|------|-----|-----------|------|
| 氯沙坦 | 160 | 13.29 | 6.10 |
| 伊贝沙坦 | 160 | 14.87 | 5.84 |

2. 如何用 SAS 实现等效性检验

【分析与解答】 该资料属于成组设计一元定量资料, 分析目的是评价两种药物的降压效果是否等效, 应该采用双单侧检验进行等效性检验, 设定等效性界限为 $L = -5\text{mmHg}$, $U = 5\text{mmHg}$ 。应用 SAS 进行分析, 程序如下, 程序名为 SAS-TJFX8_9.SAS。

```

data SASTJFX8_9;
n1 = 160; n2 = 160;
mean1 = 13.29; mean2 = 14.87;
s1 = 6.10; s2 = 5.84;
L = -5; U = 5;
ss1 = s1 ** 2 * (n1 - 1);
ss2 = s2 ** 2 * (n2 - 1);
sc = (ss1 + ss2) / (n1 + n2 - 2);
se = sqrt(sc * (1/n1 + 1/n2));

t1 = (mean1 - mean2) - L / se; /* 第1步 */
t2 = (U - (mean1 - mean2)) / se; /* 第2步 */
p1 = 1 - probt(t1, n1 + n2 - 2); /* 第3步 */
p2 = 1 - probt(t2, n1 + n2 - 2); /* 第4步 */
ods html;
proc print;
var t1 p1 t2 p2;
run;
ods html close;

```

【程序说明】 第1步对 $H_{0(1)}$, 即等效性界值的下限 L 进行假设检验, 计算 t_1 ; 第2步对 $H_{0(2)}$, 即等效性界值的上限 U 进行假设检验, 计算 t_2 ; 第3步计算 t_1 对应的 t 分布右侧的累计概率; 第4步计算 t_2 对应的 t 分布右侧的累计概率。

【主要输出结果及解释】

| Obs | t1 | p1 | t2 | p2 |
|-----|---------|------------|---------|----|
| 1 | 5.12264 | .000000262 | 9.85584 | 0 |

【统计与专业结论】 $t_1 = 5.12264$, $p_1 = 0.000000262$, 按照 $\alpha = 0.025$, 拒绝 $H_{0(1)}$, 接受

| 程序 | 说明 |
|---------------------------------|------------------------------|
| Ods html; | |
| PROC SORT; BY GROUP; RUN; | 对各组进行正态性检验 |
| PROC UNIVARIATE NORMAL; VAR x; | 调用单变量过程, 进行正态性检验, 选项 NOR- |
| BY GROUP; RUN; | MAL 表示进行正态性检验 |
| PROC GLM; CLASS GROUP; | 调用 GLM 过程进行方差分析, MODEL 语句中的选 |
| MODEL X = GROUP/SS3; | 项 SS3 表示只输出第Ⅲ型方差分析的结果, MEANS |
| MEANS GROUP/HOVTTEST SNK ; RUN; | 语句中的选项 HOVTTEST 表示进行方差齐性检验 |
| Ods html close; | 选项 SNK 使用 snk 法(q 检验)进行两两比较 |

8.6.5 主要分析结果及解释

GROUP = GROUP1

正态性检验

| 检验 | 统计量 | P 值 |
|---------------|-----------|---------------|
| Shapiro- Wilk | W 0.92112 | Pr < W 0.3664 |

GROUP = GROUP2

正态性检验

| 检验 | 统计量 | P 值 |
|---------------|------------|---------------|
| Shapiro- Wilk | W 0.955358 | Pr < W 0.7319 |

GROUP = GROUP3

正态性检验

| 检验 | 统计量 | P 值 |
|---------------|------------|---------------|
| Shapiro- Wilk | W 0.894252 | Pr < W 0.1892 |

GROUP = GROUP4

正态性检验

| 检验 | 统计量 | P 值 |
|---------------|------------|---------------|
| Shapiro- Wilk | W 0.884065 | Pr < W 0.1452 |

The GLM Procedure

Levene's Test for Homogeneity of x Variance

ANOVA of Squared Deviations from Group Means

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|--------|----|----------------|-------------|---------|--------|
| GROUP | 3 | 822.0627 | 29.3542 | 0.68 | 0.5705 |
| Error | 36 | 1555.8 | 43.2173 | | |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|--------|----|-------------|-------------|---------|---------|
| GROUP | 3 | 323.4750000 | 107.8250000 | 17.52 | < .0001 |

Means with the same letter
are not significantly different.

| SNK Grouping | Mean | N | GROUP |
|--------------|--------|----|--------|
| A | 64.300 | 10 | GROUP1 |
| A | | | |

| | | | |
|---|---------|----|--------|
| A | 63.400 | 10 | GROUP3 |
| B | 522.300 | 10 | GROUP2 |
| B | | | |
| B | 522.100 | 10 | GROUP4 |

首先查验正态性和方差齐性检验的结果：可知 4 个组均符合正态性，W 值分别为：0.92112、0.955358、0.894252、0.884065，P 值分别为：0.3664、0.7319、0.1892、0.1452；方差齐性检验的结果：F=0.68，P=0.5705>0.05，说明该定量资料满足方差齐性要求，故选用参数检验法(单因素 4 水平设计定量资料的方差分析)。由方差分析的结果可以得知：F=17.52，P<0.0001，可认为 4 种品系(分别为 GROUP1、GROUP2、GROUP3、GROUP4)小麦株高的总体均值之间的差别有统计学意义。SNK 法(即 q 检验)进行两两比较的结果可以看出，1 组与 3 组、2 组与 4 组平均值之间的差别没有统计学意义，1 组与 2 组、1 组与 4 组、2 组与 3 组、3 组与 4 组平均值之间的差别有统计学意义。

专业结论：津丰小麦 4 个品系株高总体均值之间存在差别，品系 0-3-1 与品系 0-3-3、品系 0-3-2 与品系 0-3-4 平均株高分别接近相等；品系 0-3-1 与品系 0-3-2、品系 0-3-1 与品系 0-3-4、品系 0-3-2 与品系 0-3-3、品系 0-3-3 与品系 0-3-4 平均株高不等。

8.7 单因素 k(k≥3) 水平设计定量资料一元协方差分析

8.7.1 问题与数据

【例 8-11】 研究 3 种饲料对大白鼠体重增加的影响，将体重接近、鼠龄相差较小的 24 只大白鼠随机分成 3 组，每组 8 只，喂以 3 种不同饲料，各组每只鼠在试验期间内的平均进食量与体重增加量如表 8-6 所示。试分析 3 组动物体重增加量的均数间差别有无统计学意义。

表 8-6 三组大白鼠进食量 x 与体重增加量 y 试验结果(g)

| 编号 | 进食量与增重(g) | | 编号 | 进食量与增重(g) | | 编号 | 进食量与增重(g) | |
|----|-----------|------|----|-----------|------|----|-----------|------|
| | 1 组: x | y | | 2 组: x | y | | 3 组: x | y |
| 1 | 306.9 | 45.0 | 9 | 302.4 | 50.3 | 17 | 310.3 | 61.2 |
| 2 | 256.9 | 27.3 | 10 | 260.3 | 33.4 | 18 | 250.5 | 43.8 |
| 3 | 204.5 | 25.4 | 11 | 214.8 | 36.7 | 19 | 210.4 | 39.0 |
| 4 | 272.4 | 48.0 | 12 | 278.9 | 51.2 | 20 | 275.3 | 51.5 |
| 5 | 340.2 | 56.7 | 13 | 340.9 | 58.2 | 21 | 335.1 | 66.4 |
| 6 | 198.2 | 9.2 | 14 | 199.0 | 22.5 | 22 | 199.2 | 10.8 |
| 7 | 262.2 | 28.5 | 15 | 260.5 | 27.6 | 23 | 263.3 | 25.7 |
| 8 | 247.8 | 37.1 | 16 | 240.8 | 41.0 | 24 | 245.0 | 50.9 |

8.7.2 对数据结构的分析

该例涉及一个因素(饲料种类)，该因素具有 3 个水平(即 3 种不同饲料)，观测指标为“增重”，“进食量”不是最终的观测指标，但是作为一个重要的非试验因素可能会干扰和影响“增重”，该资料类型属于单因素 3 水平一元定量资料。

8.7.3 分析目的与统计分析方法的选择

试验中，试验因素有时会受到某个重要的定量的非试验因素的影响。为了消除这种定量非试验因素对定量观测结果的影响和干扰，经常采用协方差分析。本例为消除“进食量”对“增重”的影响，可使用协方差分析。应用协方差分析的前提条件：其一，要求各组定量资料(主要指观测结果)来自方差相等的正态总体；其二，各组的总体回归斜率要相等，且不等于零。选用 glm 过程可对此定量资料进行分析。

8.7.4 SAS 程序中重要内容的说明

例 8-11 资料的程序名为 SASTJFX8_11.SAS。

| 程序 | 说明 |
|------------------------------------|------------------------------|
| DATA SASTJFX8_11; | 建立数据集 |
| DO g=1 to 8; DO a=1 to 3; | |
| INPUT X Y @@ ; OUTPUT; END;END; | 输入数据 |
| CARDS; | |
| 306.9 45.0 302.4 50.3 310.3 61.22 | |
| 56.9 27.3 260.3 33.4 250.5 43.8 | |
| 204.5 25.4 214.8 36.7 210.4 39.0 | |
| 272.4 48.0 278.9 51.2 275.3 51.5 | |
| 340.2 56.7 340.9 58.2 335.1 66.4 | |
| 198.2 9.2 199.0 22.5 199.2 10.8 | |
| 262.2 28.5 260.5 27.6 263.3 25.7 | |
| 247.8 37.1 240.8 41.0 245.0 50.9 | |
| ; | |
| RUN; | |
| Ods html;PROC SORT; BY A; RUN; | 对变量 y 进行正态性检验 |
| PROC UNIVARIATE NORMAL; | |
| VAR Y; BY A; RUN; | 调用 GLM 过程对变量 y 进行方差齐性检验 |
| PROC GLM; | |
| CLASS A; MODEL Y=A; | |
| MEANS A/HOVTTEST; RUN; | 调用 GLM 过程检验回归斜率是否相等 |
| PROC GLM; | |
| CLASS A; MODEL Y=X A X*A/SS3; RUN; | 调用 GLM 过程进行协方差分析和两两比较， |
| PROC GLM; | MODEL 语句中的选项 SOLUTION 要求输出回归 |
| CLASS A;MODEL Y=X A/SOLUTION SS3; | 系数的估计值，LSMEANS 语句中的选项 STDERR |
| LSMEANS A /STDERR PDIF; RUN; | 要求输出修正均数的标准误，选项 PDIF 要求输 |
| Ods html close; | 出修正均数两两比较的 P 值 |

8.7.5 主要分析结果及解释

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|--------|----|-------------|-------------|---------|--------|
| X | 1 | 3458.388979 | 3458.388979 | 41.38 | <.0001 |
| a | 2 | 63.479898 | 31.739949 | 0.38 | 0.6894 |
| X * a | 2 | 71.042449 | 35.521225 | 0.42 | 0.6602 |
| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
| X | 1 | 3467.987594 | 3467.987594 | 44.02 | <.0001 |
| a | 2 | 327.244289 | 163.622145 | 2.08 | 0.1515 |

Least Squares Means for effect a
Pr > |t| for H0: LSMean(i) = LSMean(j)

| Dependent Variable: Y | | | |
|-----------------------|--------|--------|--------|
| i/j | 1 | 2 | 3 |
| 1 | | 0.2578 | 0.0558 |
| 2 | 0.2578 | | 0.3968 |
| 3 | 0.0558 | 0.3968 | |

首先,应查验资料是否符合正态性与方差齐性,可知均符合参数检验前提条件,由于具体输出结果较多,此处从略。然后,检查各组斜率是否相等,因 $x * a$ 差别无统计学意义 ($F = 0.42, P = 0.6602 > 0.05$),故可认为三个饲料组斜率是相等的,因此该定量资料满足一元协方差分析的前提条件。分析结果:进食量 x 对 y 的影响有统计学意义 ($F = 44.02, P < 0.0001$),三个饲料组体重平均增加量之间的差别无统计学意义 ($F = 2.08, P = 0.1515$)。两两比较的结果也表明任意两组之间的差别都没有统计学意义。综上所述,尚不能认为三个饲料组动物体重增加量之间存在差别,故三种饲料对动物体重增加量的影响相同。

8.8 单因素 $k(k \geq 3)$ 水平设计一元定量资料 Welch 近似方差分析和 Kruskal-Wallis 秩和检验及两两比较

8.8.1 问题与数据

【例 8-12】某地检测大气中 SO_2 的浓度 ($\mu g/m^3$),按不同功能区设置采样点,结果如下所示。问:各功能区有无差别?

对照区: 10、39、45、48、50、52、44、47

工业区: 967、665、709、802、851、872、911、694

商业区: 219、541、634、669、677、610、570、321

居民区: 352、495、511、238、502、459、474、442

8.8.2 对数据结构的分析

该例涉及一个因素(地区),该因素具有 4 个水平,观测指标为“ SO_2 的浓度”,因此该资料类型属于单因素 4 水平一元定量资料。

8.8.3 分析目的与统计分析方法的选择

对单因素多水平 ($k \geq 3$) 设计定量资料进行方差分析时,定量资料应满足独立性、正态性、方差齐性。若定量资料不满足参数检验的前提条件,则可选择秩和检验进行分析。可调用 NPAR1WAY 过程达到秩和检验分析目的。若定量资料满足独立性和正态性,仅不满足方差齐性,也可采用 Welch 法进行近似的单因素多水平设计定量资料方差分析,此时的关键词句为:

```
means 因素名/hovtest welch;
```

8.8.4 SAS 程序中重要内容的说明

例 8-12 资料的程序名为 SASTJFX8_12.SAS。

| 程序 | 说明 |
|---------------------------------------|-----------------------------|
| DATA SASTJFX8_12; | 建立数据集 |
| INPUT GROUP \$; DO i =1 TO 8; | |
| INPUT x @@ ;OUTPUT; END; | 以下是输入数据 |
| CARDS; | |
| GROUP1 | |
| 10 39 45 48 50 52 44 47 | |
| GROUP2 | |
| 967 665 709 802 851 872 911 694 | |
| GROUP3 | |
| 219 541 634 669 677 610 570 321 | |
| GROUP4 | |
| 352 495 511 238 502 459 474 442 | |
| ; | |
| RUN; | |
| Ods html; | |
| PROC SORT; BY GROUP; RUN; | 对数据排序, |
| PROC UNIVARIATE NORMAL; | 对各组进行正态性检验 |
| VAR x; BY GROUP; RUN; | |
| PROC GLM; | 调用 GLM 过程 |
| CLASS GROUP; MODEL X = GROUP/SS3; | |
| MEANS GROUP/HOVTEST WELCH | 进行方差齐性检验, WELCH 要求 |
| SNK ;RUN; | 进行近似方差齐性检验 |
| PROC NPAR1WAY WILCOXON; | 调用非参数检验过程 |
| CLASS GROUP; VAR x; RUN; QUIT; | |
| PROC RANK OUT = BBB; VAR X; RANKS RX; | 调用 RANK 过程对 3 组数据排序, 排序后生成新 |
| PROC ANOVA DATA = BBB; | 变量 RX, 对排序后的数据调用 ANOVA 过程进行 |
| CLASS GROUP; MODEL RX = GROUP; | 两两比较 |
| MEANS GROUP/SNK; RUN; | |
| Ods html close; | |

8.8.5 主要分析结果及解释

GROUP = GROUP1

正态性检验

| 检验 | 统计量 | P 值 |
|---------------|------------|---------------|
| Shapiro- Wilk | W 0.694328 | Pr < W 0.0019 |

GROUP = GROUP2

正态性检验

| 检验 | 统计量 | P 值 |
|---------------|------------|---------------|
| Shapiro- Wilk | W 0.938847 | Pr < W 0.5998 |

GROUP = GROUP3

正态性检验

| 检验 | 统计量 | P 值 |
|---------------|------------|---------------|
| Shapiro- Wilk | W 0.822003 | Pr < W 0.0490 |

GROUP = GROUP4

正态性检验

| 检验 | 统计量 | P 值 |
|---------------|------------|---------------|
| Shapiro- Wilk | W 0.804174 | Pr < W 0.0317 |

以上是对 4 个组定量资料进行正态性检验的结果, 仅第 2 组符合要求。

| Levene's Test for Homogeneity of x Variance | | | | | |
|--|----|----------------|-------------|---------|--------|
| ANOVA of Squared Deviations from Group Means | | | | | |
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| GROUP | 3 | 2.6029E9 | 8.6762E8 | 2.72 | 0.0635 |
| Error | 28 | 8.9386E9 | 3.1923E8 | | |

以上是对 4 个组定量资料进行方差齐性检验，满足了方差齐性要求。

| Welch's ANOVA for x | | | |
|---------------------|---------|---------|---------|
| Source | DF | F Value | Pr > F |
| GROUP | 3.0000 | 171.48 | < .0001 |
| Error | 11.9786 | | |

以上是对 4 个组定量资料进行近似方差分析的结果，此结果仅当全部定量资料都满足正态分布要求但不满足方差齐性要求时，用于对 4 个组定量资料均值之间差别是否具有统计学意义下结论。

| Wilcoxon Scores (Rank Sums) for Variable x | | | | | |
|--|---|---------------|-------------------|------------------|------------|
| Classified by Variable GROUP | | | | | |
| GROUP | N | Sum of Scores | Expected Under H0 | Std Dev Under H0 | Mean Score |
| GROUP1 | 8 | 36.0 | 132.0 | 22.978251 | 4.5000 |
| GROUP2 | 8 | 226.0 | 132.0 | 22.978251 | 28.2500 |
| GROUP3 | 8 | 151.0 | 132.0 | 22.978251 | 18.8750 |
| GROUP4 | 8 | 115.0 | 132.0 | 22.978251 | 14.3750 |

以上是和秩和检验对应的部分统计量的数值，最后一列平均得分就是平均秩，可用于下专业结论。

| Kruskal-Wallis Test | |
|---------------------|---------|
| Chi-Square | 26.5653 |
| DF | 3 |
| Pr > Chi-Square | < .0001 |

以上是对单因素 4 水平设计一元定量资料秩和检验的结果。

| The ANOVA Procedure | | | |
|---|---------|---|--------|
| Student-Newman-Keuls Test for RX | | | |
| Means with the same letter are not significantly different. | | | |
| SNK Grouping | Mean | N | GROUP |
| A | 222.250 | 8 | GROUP2 |
| B | 122.875 | 8 | GROUP3 |
| C | 14.375 | 8 | GROUP4 |
| D | 4.500 | 8 | GROUP1 |

【输出结果解释与结论】 首先查验正态性和方差齐性检验的结果：可知第 1 组、第 3 组、第 4 组不符合正态性，P 值分别为 0.0019、0.0490、0.0317，故选用非参数检验方法（对于单因

素多水平设计一元定量资料非参数检验方法,一般选用 Kruskal-Wallis 检验)。由 Kruskal-Wallis 检验的结果可以得知: $\chi^2 = 26.5653$, $P < 0.001$, 可认为 4 个组平均秩之间差别有统计学意义。由具体两两比较的结果可以看出, 4 个平均秩中任何 2 个平均秩之间的差别均具有统计学意义。

专业结论: 对照区、工业区、商业区、居民区大气中 SO_2 的平均浓度不同, SO_2 的平均浓度由高至低依次为工业区、商业区、居民区、对照区。

8.9 本章小结

本章全面介绍了 4 种单因素设计(即单组设计、配对设计、成组设计和单因素多水平设计)一元定量资料的统计分析方法及 SAS 实现方法, 包括参数检验法(t 检验和方差分析)与非参数检验法(符号检验、符号秩和检验、Wilcoxon 秩和检验及 Kruskal-Wallis 秩和检验)。值得注意的是, 当选用参数检验法时, 定量资料必须满足特定的前提条件, 对于单组设计和配对设计一元定量资料而言, 应满足独立性和正态性两个前提条件; 对成组设计和单因素多水平设计一元定量资料而言, 应满足独立性、正态性和方差齐性三个前提条件。独立性需要使用者根据具体情况去判定其是否满足, 而检查定量资料是否满足正态性可用 UNIVARIATE 过程中的选项 NORMAL 来实现(看其中的 W 检验的结果即可); 成组设计定量资料的方差齐性检验包含在 TTEST 过程输出结果之中; 而单因素多水平设计一元定量资料的方差齐性检验需要在 GLM 过程步的 MEANS 语句中增加选项 HOVTEST 来实现。

(郭晋 胡良平)

第9章 单因素设计一元生存资料差异性分析

在医学随访研究中,有时观察结果并非能在短期内确定,需做长期随访观察。例如,对一些慢性病或恶性肿瘤的预后及远期疗效观察等。这种情况下,原有的疗效指标,如有效率、治愈率等难以适用,因为评价某种疗法对这些疾病的效果,不仅要看是否出现了感兴趣的终点事件(terminal event),还要考虑达到终点所经历的时间长短。

本章主要介绍生存率估计的 Kaplan-Meier 法和寿命表法、生存曲线比较的 log-rank 检验以及 SAS 软件中的 LIFETEST 过程及其应用。

9.1 单因素设计一元生存资料分析简介

生存分析(survival analysis)是将终点事件的出现与否和达到终点所经历的时间结合起来的一种统计分析方法,其主要特点就是考虑了每个观察对象达到终点所经历的时间长短。终点事件不限于死亡,可以是疾病的发生、一种处理(治疗)的反应、疾病的复发等。生存分析可用于生存曲线估计、生存曲线比较、影响因素分析和生存预测。

生存分析有一套完整的方法:统计描述(包括求生存时间的分位数、中位生存期、平均数、生存函数的估计、判断生存时间分布的图示法);非参数检验(检验分组变量各水平所对应的生存曲线是否一致);Cox 模型(半参数)回归分析;参数模型回归分析。

本章主要介绍单因素设计一元生存资料的分析方法,包括统计描述和非参数检验。其中,研究者比较 k 条生存曲线之间是否有显著性差别时,SAS 软件提供了 3 种常用的方法:对数秩检验(log-rank test),威尔考克森检验(Wilcoxon test)和似然比检验(likelihood ratio test)。

当生存时间的分布为 Weibull 分布或属于比例风险比模型时,log-rank 检验效率较高;当生存时间的分布为对数正态分布时,Wilcoxon 检验效率较高;似然比检验是建立在指数分布模型上的,故当资料偏离此模型时,其结果不如前两种检验方法稳健。

9.2 生存资料统计描述

9.2.1 问题与数据

【例9-1】某医师收集 11 例脑瘤患者甲疗法治疗的生存时间(周)。试估计治疗后不同时间的生存率、生存曲线及中位生存期。

甲疗法组 5 7⁺ 13 13 23 30 30⁺ 38 42 42 45⁺

【例9-2】收集 374 名某恶性肿瘤病人随访资料,取时间区间均为 1 年,整理结果见表 9-1,试估计确诊后不同年数的生存率、生存曲线、中位生存期等。

表 9-1 某恶性肿瘤病人随访资料与生存率计算过程

| 确诊后年数 (t_i) | 期内死亡数 (d_i) | 期内截尾数 (c_i) | 期初病例数 (n_i) | 确诊后年数 (t_i) | 期内死亡数 (d_i) | 期内截尾数 (c_i) | 期初病例数 (n_i) |
|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| 0 ~ | 90 | 0 | 374 | 5 ~ | 7 | 9 | 95 |
| 1 ~ | 76 | 0 | 284 | 6 ~ | 4 | 9 | 79 |
| 2 ~ | 51 | 0 | 208 | 7 ~ | 1 | 3 | 66 |
| 3 ~ | 25 | 12 | 157 | 8 ~ | 3 | 5 | 62 |
| 4 ~ | 20 | 5 | 120 | 9 ~ 10 | 2 | 5 | 54 |

注：生存时间长于 10 年者 47 例。

9.2.2 对数据结构的分析

对于例 9-1 资料，研究者选取 11 名同质的脑瘤患者，观察他们采用甲疗法治疗后的生存时间。从设计类型来说，本例应为单组设计。因生存时间是定量的观测指标，故资料类型为单组设计一元定量资料。当然，由于部分受试对象的观测数据是删失的，所以此资料是单组设计一元生存资料。

例 9-2 资料与例 9-1 相似，设计类型为单组设计，资料本质为单组设计一元生存资料。

9.2.3 分析目的与统计分析方法的选择

非参数法估计生存率有乘积极限法 (Product-limit method, 简称 PL 法) 和寿命表法 (life table method, actuarial method)，其中乘积极限法又称为 Kaplan-Meier 法 (简称 KM 法)。前者适用于小样本或大样本未分组资料，后者适用于观察例数较多的分组资料。

对例 9-1 资料，甲疗法组的生存数据为含有截尾数据的小样本未分组资料，宜采用 Kaplan-Meier 法估计生存率和生存曲线。

对例 9-2 资料，该恶性肿瘤随访资料观察例数较多，且根据随访时间和随访结果编制成频数分布表的形式，处理此类资料宜采用寿命表法估计生存率。

9.2.4 SAS 程序

分析例 9-1 资料，采用 Kaplan-Meier 法，程序名为 SASTJFX9_1.SAS。

| | |
|--|---|
| <pre>DATA SASTJFX9_1; INPUT t status @@ ; CARDS; 5 1 7 0 13 1 13 23 1 30 1 30 0 38 42 1 42 1 45 0 ; run;</pre> | <pre>ods html; PROC LIFETEST METHOD=PL PLOTS=(S); TIME t* status(0); RUN; ods html close;</pre> |
|--|---|

数据步中变量 t、status 分别表示生存时间和生存结局，其中 status = 1 表示某终点事件发生，如死亡；status = 0 表示截尾。

LIFETEST 过程可对生存资料进行 KM 法或寿命表法估计，并对两组或多组生存率做组间比较，包括 log-rank 检验、Wilcoxon 检验和似然比检验，其中似然比检验基于指数分布。

PROC LIFETEST 语句中的选项如下。

(1) METHOD = PL|KM|LT|LIFE|ACT, 指定生存率估计方法。PL 和 KM 表示乘积极限法或 KM 法 (缺省值)；LT、LIFE 和 ACT 表示寿命表法。当使用寿命表法时，SAS 可自动生成生存时间的分组区间，也可用“WIDTH = ”人为指定区间宽度，或用“INTERVALS = (a to b by

c)”，a、b、c 分别表示区间的初值、终值和组区间的宽度。

PLOTS = ()，要求绘图。括号内可填写的内容有 S、LS、LLS、H，若同时填多项，各项之间用逗号隔开，H 选项仅在使用寿命表法时有效。各选项的具体含义如下：SURVIVAL(或 S)表示生存率 $\hat{S}(t)$ 对生存时间 t 的图形，即生存曲线；LOGSURV(或 LS)表示 $-\log\hat{S}(t)$ 对 t 的图形，用于判断生存时间是否服从指数分布。若呈指数分布，则图形应呈过原点的直线；LOGLOGS(或 LLS)表示 $\log[-\log\hat{S}(t)]$ 对 $\log t$ 的图形，用于判断生存时间是否服从 Weibull 分布。若呈 Weibull 分布，则图形应呈直线；HAZARD(或 H)表示累积风险 $\hat{H}(t)$ 对 t 的图形。只有当使用寿命表法时，才出现 HAZARD 图形。

TIME 语句为 LIFETEST 过程的必须语句，设置生存时间变量和生存结局变量，括号内为截尾变量的标示值。

分析例 9-2 资料，采用寿命表法，程序名为 SASTJFX9_2.SAS。

```
DATA SASTJFX9_2;
INPUT t status number @@ ;
CARDS;
0.5 1 90 0.5 0 0 1.5 1
1.5 0 0 2.5 1 51 2.5 0
3.5 1 25 3.5 0 12 4.5 1
4.5 0 5 5.5 1 7 5.5 0
6.5 1 4 6.5 0 9 7.5 1
7.5 0 3 8.5 1 3 8.5 0
9.5 1 2 9.5 0 5 10.5 1
10.5 0 47
;
run;

ods html;
PROC LIFETEST METHOD=LIFE WIDTH=1;
TIME t*status(0);
FREQ number;
RUN;
ods html close;
```

数据步中变量 t、status 和 number 分别表示每个时间区间的中点、生存结局(死亡 = 1，截尾 = 0)及其频数。过程步中指定生存率的估计方法为寿命表法，时间区间宽度为 1。同时，用 FREQ 语句指定频数变量。

9.2.5 主要分析结果及解释

以下是 SAS 软件输出的关于例 9-1 的分析结果。

| Product-Limit Survival Estimates | | | | | |
|----------------------------------|----------|---------|-------------------------|---------------|-------------|
| t | Survival | Failure | Survival Standard Error | Number Failed | Number Left |
| 0 | 1.0000 | 0 | 0 | 0 | 11 |
| 5 | 0.9091 | 0.0909 | 0.0867 | 1 | 10 |
| 7 * | . | . | . | 1 | 9 |
| 13 | . | . | . | 2 | 8 |
| 13 | 0.7071 | 0.2929 | 0.1429 | 3 | 7 |
| 23 | 0.6061 | 0.3939 | 0.1541 | 4 | 6 |
| 30 | 0.5051 | 0.4949 | 0.1581 | 5 | 5 |
| 30 * | . | . | . | 5 | 4 |
| 38 | 0.3788 | 0.6212 | 0.1613 | 6 | 3 |
| 42 | . | . | . | 7 | 2 |
| 42 | 0.1263 | 0.8737 | 0.1163 | 8 | 1 |
| 45 * | . | . | . | 8 | 0 |

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable t

Quartile Estimates

| | Point | 95% Confidence Interval | |
|---------|----------|-------------------------|---------|
| Percent | Estimate | [Lower | Upper) |
| 75 | 42.0000 | 30.0000 | . |
| 50 | 38.0000 | 13.0000 | 42.0000 |
| 25 | 13.0000 | 5.0000 | 38.0000 |
| | Mean | Standard Error | |
| | 29.1414 | 4.5151 | |

NOTE: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Summary of the Number of Censored and Uncensored Values

| Total | Failed | Censored | Percent Censored |
|-------|--------|----------|------------------|
| 11 | 8 | 3 | 27.27 |

输出的第一部分是甲疗法组生存率的 KM 法估计结果。统计量包括各生存时间点对应的生存率 (Survival)、死亡率 (Failure)、生存率标准误 (Survival Standard Error)、累积死亡数 (Number Failed) 和期末例数 (Number Left)。

结果表明, 采用甲疗法治疗, 5 周生存率为 90.91%, 13 周生存率为 70.71%, 其他依此类推。注意: (1) 生存时间一栏标有“*”者为截尾生存时间, 其生存率和生存率标准误用“.”表示, 实际上该截尾生存时间的生存率和生存率标准误与前一个完全生存时间对应数值相同, 如 7 周生存率为 90.91%。(2) 两个完全生存时间之间任意时间的生存率均等于前一个完全生存时间的生存率, 如 6 个月 (即 24 周) 生存率为 60.61%。

第二部分输出生存时间的分位数, 包括 75%、50% 和 25% 分位数的点估计及 95% 置信区间。50% 分位数即中位生存期, 为 38 周, 95% 置信区间 19~42(周)。

最后一部分概括截尾和非截尾例数。本资料总例数为 11, 其中死亡 8 例, 截尾 3 例, 截尾百分比为 27%。

Kaplan-Meier 生存曲线为阶梯形曲线, 见图 9-1。

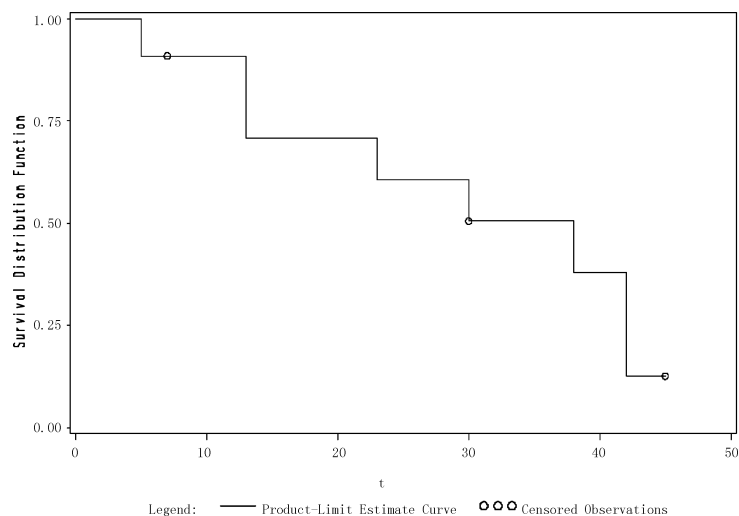


图 9-1 甲疗法治疗脑瘤患者 Kaplan-Meier 生存曲线

以下是 SAS 软件输出的关于例 9-2 的分析结果。

| Interval [Lower Upper) | | Number Failed | Number Censored | Effective Sample Size | Conditional Probability of Failure | Conditional Probability Standard Error | Survival |
|---------------------------------|----|------------------|--------------------|--------------------------|--|--|----------|
| 0 | 1 | 90 | 0 | 374.0 | 0.2406 | 0.0221 | 1.0000 |
| 1 | 2 | 76 | 0 | 284.0 | 0.2676 | 0.0263 | 0.7594 |
| 2 | 3 | 51 | 0 | 208.0 | 0.2452 | 0.0298 | 0.5561 |
| 3 | 4 | 25 | 12 | 151.0 | 0.1656 | 0.0302 | 0.4198 |
| 4 | 5 | 20 | 5 | 117.5 | 0.1702 | 0.0347 | 0.3503 |
| 5 | 6 | 7 | 9 | 90.5 | 0.0773 | 0.0281 | 0.2907 |
| 6 | 7 | 4 | 9 | 74.5 | 0.0537 | 0.0261 | 0.2682 |
| 7 | 8 | 1 | 3 | 64.5 | 0.0155 | 0.0154 | 0.2538 |
| 8 | 9 | 3 | 5 | 59.5 | 0.0504 | 0.0284 | 0.2498 |
| 9 | 10 | 2 | 5 | 51.5 | 0.0388 | 0.0269 | 0.2372 |
| 10 | . | 0 | 47 | 23.5 | 0 | 0 | 0.2280 |

| Evaluated at the Midpoint of the Interval | | | | | | | |
|---|----------|----------|----------|---------|-------------------|----------|-------------------|
| Failure | Survival | Median | Median | | | | |
| | Standard | Residual | Standard | PDF | | Hazard | |
| | Error | Lifetime | Error | PDF | Standard Error | Hazard | Standard Error |
| 0 | 0 | 2.4118 | 0.1896 | 0.2406 | 0.0221 | 0.273556 | 0.028564 |
| 0.2406 | 0.0221 | 2.5771 | 0.3242 | 0.2032 | 0.0208 | 0.308943 | 0.035013 |
| 0.4439 | 0.0257 | 3.5599 | 0.8576 | 0.1364 | 0.0177 | 0.279452 | 0.038747 |
| 0.5802 | 0.0255 | . | . | 0.0695 | 0.0134 | 0.180505 | 0.035954 |
| 0.6497 | 0.0248 | . | . | 0.0596 | 0.0129 | 0.186047 | 0.041421 |
| 0.7093 | 0.0239 | . | . | 0.0225 | 0.00837 | 0.08046 | 0.030386 |
| 0.7318 | 0.0235 | . | . | 0.0144 | 0.00712 | 0.055172 | 0.027576 |
| 0.7462 | 0.0233 | . | . | 0.00393 | 0.00392 | 0.015625 | 0.015625 |
| 0.7502 | 0.0233 | . | . | 0.0126 | 0.00718 | 0.051724 | 0.029853 |
| 0.7628 | 0.0232 | . | . | 0.00921 | 0.00645 | 0.039604 | 0.027999 |
| 0.7720 | 0.0232 | . | . | . | . | . | . |

Summary of the Number of Censored and Uncensored Values

| Total | Failed | Censored | Percent Censored |
|-------|--------|----------|------------------|
| 374 | 279 | 95 | 25.40 |

NOTE: There were 4 observations with missing values, negative time values or frequency values less than 1.

SAS 输出统计量包括各生存时间区间内 (Interval) 的死亡数 (Number Failed)、截尾数 (Number Censored)、期初有效例数 (Effective Sample Size)、条件死亡概率 (Conditional Probability of Failure)、条件死亡概率标准误 (Conditional Probability Standard Error)、区间左端点处生存率 (Survival)、死亡率 (Failure)、生存率标准误 (Survival Standard Error)、中位剩余寿命 (Median Residual Lifetime)、中位剩余寿命标准误 (Median Standard Error) 及各时间区间中点处的概率密度函数 (PDF)、概率密度函数标准误 (PDF Standard Error)、风险函数 (Hazard) 和风险函数标准误 (Hazard Standard Error)。

结果表明,该类恶性肿瘤患者1年生存率为75.94%,2年生存率为55.61%,其他依此类推。

9.3 生存曲线比较

9.3.1 问题与数据

【例9-3】沿用例9-1,某医师同时收集了9例脑瘤患者乙疗法治疗的生存时间(周)。试比较甲、乙两疗法组生存率之间的差别有无统计学意义。

乙疗法组 1 3 3 7 10 15 15 23 30

9.3.2 对数据结构的分析

研究者收集了甲、乙两种疗法治疗肿瘤患者后的生存时间,试验中涉及一个试验因素,即治疗方法,此因素有两个水平:甲疗法和乙疗法。生存时间是定量指标,所以此资料是成组设计一元生存资料。

9.3.3 分析目的与统计分析方法的选择

log-rank 检验和 Wilcoxon 检验是生存率比较的非参数方法,由于能对各组的生存曲线做整体比较,在实际工作中应用较多。其中,log-rank 检验给组间死亡的远期差别更大的权重,即对远期差异敏感;Wilcoxon 检验给组间死亡的近期差别更大的权重,即对近期差异敏感。本例两条生存曲线基本呈比例(见图9-2),选用 log-rank 检验。

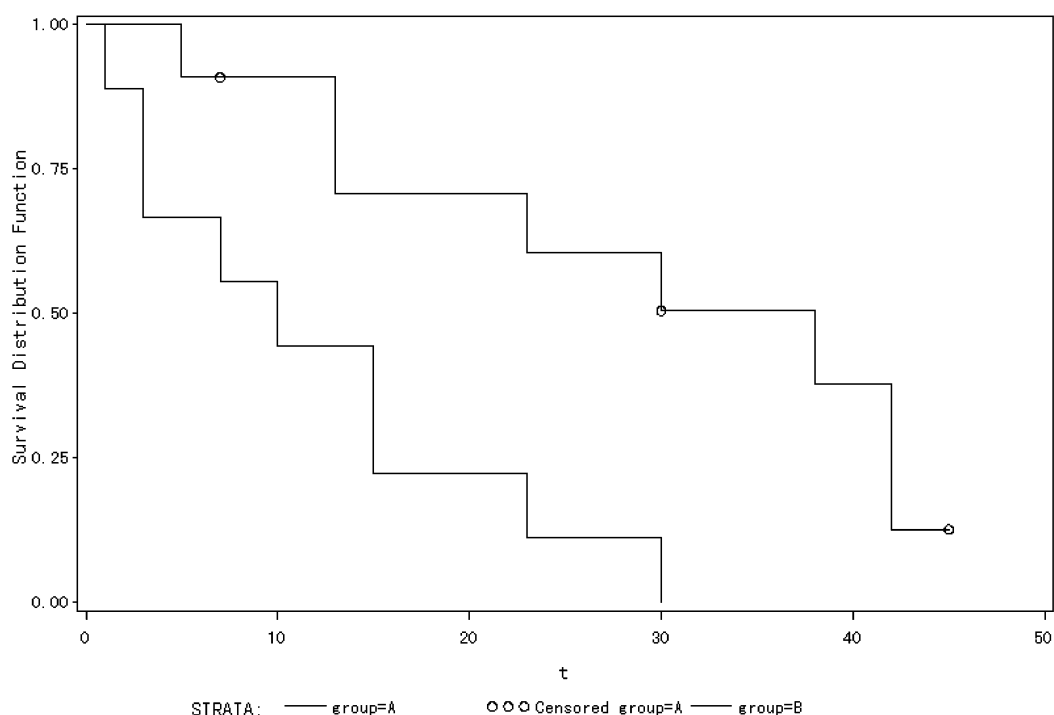


图9-2 脑瘤患者甲、乙两疗法组生存曲线

9.3.4 SAS 程序

程序名为 SASTJFX9_3.SAS。

```
DATA SASTJFX9_3;
INPUT group $ n;
DO I=1 TO n;
INPUT t status @@ ;
OUTPUT;
END;
CARDS;
A 11
5 1 7 0 13 1 13 1 23 1 30 1 30 0 38 1 42 1 42 1 45 0
B 9
1 1 3 1 3 1 7 1 10 1 15 1 15 1 23 1 30 1
; run;
ods html;
PROC LIFETEST METHOD=PL PLOTS=(S);
TIME t*status(0);
STRATA group;
RUN;
ods html close;
```

与程序 SASTJFX9_1.SAS 不同的是，该程序的 LIFETEST 过程步使用了“STRATA group;”语句，这个语句用来指定分组变量，从而实现两组或多组生存曲线间的比较。

9.3.5 主要分析结果及解释

| Summary of the Number of Censored and Uncensored Values | | | | | |
|---|-------|-------|--------|----------|------------------|
| Stratum | group | Total | Failed | Censored | Percent Censored |
| 1 | A | 11 | 8 | 3 | 27.27 |
| 2 | B | 9 | 9 | 0 | 0.00 |
| Total | | 20 | 17 | 3 | 15.00 |

| Testing Homogeneity of Survival Curves for t over Strata | | |
|--|-----------|----------|
| Rank Statistics | | |
| group | Log- Rank | Wilcoxon |
| A | -4.5500 | -62.000 |
| B | 4.5500 | 62.000 |

| Covariance Matrix for the Log- Rank Statistics | | |
|---|----------|----------|
| group | A | B |
| A | 2.71390 | -2.71390 |
| B | -2.71390 | 2.71390 |

| Covariance Matrix for the Wilcoxon Statistics | | |
|--|---|---|
| group | A | B |

| | | |
|---|----------|----------|
| A | 587.122 | -587.122 |
| B | -587.122 | 587.122 |

Test of Equality over Strata

| Test | Chi-Square | DF | Pr > Chi-Square |
|-----------|------------|----|-----------------|
| Log-Rank | 7.6283 | 1 | 0.0057 |
| Wilcoxon | 6.5472 | 1 | 0.0105 |
| -2Log(LR) | 5.0557 | 1 | 0.0245 |

生存曲线组间比较结果, 包括 log-rank 检验和 Wilcoxon 检验的秩统计量(Rank Statistics)、协方差矩阵(Covariance Matrix)、 χ^2 统计量和 P 值。

log-rank 检验结果 Chi-Square 为 7.63, 自由度为 1, P 值为 0.0057。可认为两条生存曲线不同, 甲疗法组患者的生存曲线高于乙疗法组, 即甲疗法优于乙疗法。

9.4 本章小结

(1) 生存曲线的非参数估计法有 Kaplan-Meier 法和寿命表法。Kaplan-Meier 法适用于小样本或大样本未分组资料, 寿命表法适用于观察例数较多的分组资料, 二者均利用概率乘法定理计算生存率。SAS 实现过程为 LIFETEST。

(2) log-rank 检验是两条或多条生存曲线比较的非参数方法之一, 由于该检验能对各组的生存曲线做整体比较, 在实际工作中应用较多。SAS 实现过程为 LIFETEST。

(余红梅 高辉)

第 10 章 多因素设计一元定量资料差异性分析

随着科研工作的深入发展，科研人员在实验时所考虑的实验因素或重要非实验因素越来越多，多因素实验设计也应用更加频繁，所得的统计结果说服力也更强。当实验中涉及多个实验因素，而观测指标的性质是定量的且只有一个或多个观测指标但相互之间在专业上没有联系，需逐一进行分析时，这样的资料就是多因素设计一元定量资料。

对其进行统计分析时，需把握以下几个关键点：研究者的研究目的；因素间的关系有无专业依据；资料的实验设计类型；资料是否满足参数检验的前提条件。本章着重讲述常见多因素实验设计一元定量资料方差分析、协方差分析的 SAS 实现及单因素 2 水平设计一元定量资料的 meta 分析。

10.1 随机区组设计一元定量资料方差分析与 Friedman 秩和检验

10.1.1 问题与数据

【例 10-1】 某研究者欲比较 3 种抗癌药物对小白鼠肉瘤的抑瘤疗效，先将 15 只染有肉瘤的小白鼠按体重大小配成 5 个区组，使每个区组内的 3 只小白鼠体重最接近，然后随机决定每个区组中的 3 只小白鼠接受 3 种药物的治疗，以肉瘤的重量为指标，实验结果见表 10-1。试比较不同抗癌药物对小白鼠肉瘤的抑瘤效果之间的差别是否有统计学意义。

表 10-1 不同药物作用后小白鼠肉瘤重量

| 区组 | 肉瘤重量(g) | | | |
|----|---------|------|------|------|
| | 药物种类： | A | B | C |
| 1 | | 0.82 | 0.65 | 0.51 |
| 2 | | 0.73 | 0.54 | 0.23 |
| 3 | | 0.43 | 0.34 | 0.28 |
| 4 | | 0.41 | 0.21 | 0.31 |
| 5 | | 0.68 | 0.43 | 0.24 |

10.1.2 对数据结构的分析

此实验中，实验因素为“药物种类”，它有 3 个水平，分别为 A、B、C 药。受试对象为小白鼠，观测指标为肉瘤重量。此外，还有一个区组因素，即按体重形成的区组。研究者先按照体重大小将 15 只小白鼠配成 5 个区组，然后将每个区组内的 3 只小白鼠随机分到 3 个药物组内进行实验，因此此设计应为随机区组设计。

10.1.3 分析目的与统计分析方法的选择

研究者欲比较不同抗癌药物对小白鼠肉瘤的抑瘤效果之间的差别是否有统计学意义。在资料满足参数检验的前提条件时，可采用随机区组设计定量资料方差分析，若资料不满足参数检验的前提条件，可采用 Friedman 秩和检验。

10.1.4 SAS 程序

SAS 程序名为 sastjfx10_1A.sas。

```

data sastjfx10_1A;                                /* 1 */
  do a=1 to 5;
  do b=1 to 3;
  input y @@ ;
  output;
  end;
  end;
  cards;
0.82 0.65 0.51
0.73 0.54 0.23
0.43 0.34 0.28
0.41 0.21 0.31
0.68 0.43 0.24
;
run;
proc univariate normal noprint data=
sastjfx10_1A;                                     /* 2 */
  var y;
  class a;
  output out=set1
    normal=w probn=p;
run;
proc univariate normal noprint data=
sastjfx10_1A;                                     /* 3 */
  var y;
  class b;
  output out=set2
    normal=w probn=p;
run;
ods html;
data c;                                           /* 4 */
  set set1 set2;
  file print ods=(variables
    = (a b w p));
  title 'This is the results of
    Normality test';
  put_ods_;
run;
title;
proc glm data=sastjfx10_1A; /* 5 */
  class a;
  model y=a/ss3;
  means a/hovtest;
  ods output HOVFTest=set3;
run;
proc glm data=sastjfx10_1A; /* 6 */
  class b;
  model y=b/ss3;
  means b/hovtest;
  ods output HOVFTest=set4;
run;
data d;                                           /* 7 */
  set set3 set4;
  if Source='Error' then delete;
  file print ods=(variables=
    (Source FValue ProbF));
  title 'This is the results of Homo-
    geneity of Variance test';
  put_ods_;
run;
title;
proc glm data=sastjfx10_1A; /* 8 */
  class a b;
  model y=a b/ss3;
  means b/snk;
run;
ods html close;

```

程序第 1 步通过两个循环语句建立数据集 sastjfx10_1A, 其中 a 代表区组因素, b 代表实验因素“药物种类”。第 2 步、第 3 步分别对 a 因素各水平组、b 因素各水平组进行正态性检验, 并将正态性检验结果分别输出到数据集 set1、set2 中, 正态性检验统计量记为 w, 相应概率记为 p, 以便后面调用。第 4 步合并两因素各水平正态性检验结果, 以便将其一起输出, 便于查看。第 5 步、第 6 步分别对 a 因素各水平组、b 因素各水平组进行方差齐性检验, 并将方差齐性的检验结果分别输出到数据集 set3、set4 中, 采用的方法是 SAS 默认的 levene 法。第 7 步合并两因素各水平方差齐性的检验结果, 以便将其一起输出, 便于查看。第 8 步是调用 glm 过程进行随机区组设计定量资料方差分析, 并对实验因素 b 各水平进行两两比较。

若资料不满足参数检验的前提条件, 可采用 Friedman 非参数检验方法。假设例 10-1 资料不满足参数检验的前提条件, 需进行 Friedman 非参数检验, SAS 程序见 sastjfx10_1B.sas。

```

data sastjfx10_1B;
  do a=1 to 5;
    do b=1 to 3;
      inputy @@ ;
      output;
    end;
  end;
  cards;
0.82 0.65 0.51
0.73 0.54 0.23
0.43 0.34 0.28
0.41 0.21 0.31
0.68 0.43 0.24
;
run;

ods html;
proc freq data=sastjfx10_1B;
  tables a*b*y/CMH2 scores=rank
  noprint;
run;
ods html close;

```

10.1.5 主要分析结果及解释

This is the results of Normality test

| a | b | the normality test statistic, yp-value of normality test stat, y | |
|---|---|--|--------------|
| 1 | . | 0.9968879668 | 0.8934020134 |
| 2 | . | 0.9811616954 | 0.737036565 |
| 3 | . | 0.9868421053 | 0.7804408149 |
| 4 | . | 1 | 1 |
| 5 | . | 0.9938398357 | 0.8499470899 |
| . | 1 | 0.8765323942 | 0.2938917997 |
| . | 2 | 0.992862954 | 0.9886970152 |
| . | 3 | 0.7906109476 | 0.0677941326 |

这是分别对区组因素 a 和实验因素 b 所做的正态性检验结果，第 1 列为区组因素 a 及其 5 个水平，第 2 列为实验因素 b 及其 3 个水平，第 3 列为正态性检验的统计量 W 值，第 4 列为统计量 W 对应的 P 值。由第四列可以看出，a、b 两因素各水平对应的正态性检验结果 P 值均大于 0.05，可认为资料满足正态性要求。

This is the results of Homogeneity of Variance test

| Source | F Value | Pr > F |
|--------|---------|--------|
| a | 1.76 | 0.2140 |
| b | 1.06 | 0.3782 |

这是对 a、b 两因素各水平进行方差齐性检验的结果，对应的 P 值分别为 0.2140、0.3782，均大于 0.05，可认为满足方差齐性条件。为节省篇幅，此处只列出了方差齐性检验的结果。其实，在对 a、b 两因素各水平进行方差齐性检验时，程序运行的结果中还列出了以 a 因素为分组因素进行单因素 5 水平设计定量资料方差分析的结果，列出了以 b 因素为分组因素进行单因素 3 水平设计定量资料方差分析的结果，这些结果是方差齐性检验的副产品，与随机区组设计定量资料方差分析主题无关，读者可不予关注。

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|--------|----|-------------|-------------|---------|--------|
| a | 4 | 0.22836000 | 0.05709000 | 5.98 | 0.0158 |
| b | 2 | 0.22800000 | 0.11400000 | 11.94 | 0.0040 |

这是随机区组设计一元定量资料方差分析的结果, a、b 两因素对应的 P 值分别为 0.0158、0.0040, 说明两因素各水平对观测指标的影响之间的差异都有统计学意义, 即不同体重、不同药物对肉瘤重量的影响是不完全相同的。

Student-Newman-Keuls Test for y

Means with the same letter

are not significantly different.

| SNK Grouping | Mean | N | b |
|--------------|---------|---|---|
| A | 0.61400 | 5 | 1 |
| B | 0.43400 | 5 | 2 |
| B | 0.31400 | 5 | 3 |

这是对实验因素 b 各水平组之间进行两两比较的结果, 由于“SNK Grouping”列字母不全相同, 说明因素 b 的 1 水平和 2、3 水平之间的差异有统计学意义, 而 2 水平和 3 水平之间的差异则无统计学意义。b 因素 3 个水平所对应的观测指标的均值分别为 0.614、0.434、0.314, 说明 C 药对肉瘤重量的影响大于 A 药和 B 药, 因为肉瘤重量越小, 药物的医疗效果就越好。

以下是对例 10-1 进行 Friedman 秩和检验的结果。

The FREQ Procedure

Summary Statistics for b by y

Controlling for aCochran-Mantel-Haenszel Statistics (Based on Rank Scores)

| Statistic | Alternative Hypothesis | DF | Value | Prob |
|-----------|------------------------|----|--------|--------|
| 1 | Nonzero Correlation | 1 | 8.1000 | 0.0044 |
| 2 | Row Mean Scores Differ | 2 | 8.4000 | 0.0150 |

对随机区组设计定量资料进行 Friedman 秩和检验时, 可采用 CMH 检验进行统计分析, 选择 scores = rank 选项, 那么 CMH 检验计算出来的行平均得分差异统计量就等于 Friedman 秩和检验的结果。注意: CMH 检验计算时以区组因素 a 为分层变量, 以实验因素 b 为行变量, 以结果变量 y 为列变量, 构筑一个三维列联表, 所以在 tables 语句后一定要注意三个变量的写作顺序——区组因素 * 实验因素 * 结果变量。查看上表中的行平均得分差异统计量, 其值为 8.40, 对应的 P 值为 0.0150, 说明 b 因素各水平对观测指标平均值的影响之间的差异有统计学意义。

10.2 双因素无重复实验设计一元定量资料方差分析

10.2.1 问题与数据

【例 10-2】某医生欲研究回心草各单体成分对实验性心肌缺血血流动力学的影响, 选取健康新西兰家兔若干只, 体重(2.0 ± 0.3)kg, 雌雄不计, 将其随机分成 9 组: 胡椒碱高剂量组 (100nmol/L)、胡椒碱中剂量组 (10nmol/L)、胡椒碱低剂量组 (1nmol/L)、胡椒酸甲酯高剂量组 (100nmol/L)、胡椒酸甲酯中剂量组 (10nmol/L)、胡椒酸甲酯低剂量组 (1nmol/L)、咖啡酸甲酯高剂量组 (100nmol/L)、咖啡酸甲酯中剂量组 (10nmol/L)、咖啡酸甲酯低剂量组 (1nmol/L)。所有家兔处死后, 建造缺血缺氧的离体心脏模型, 给予各实验组相应种类及浓度的药物进行实验, 记录各组实验家兔血流动力学指标的平均值, 结果见表 10-2。试分析回心草的不同单体成分及给药剂量对冠脉流量的影响之间的差别是否有统计学意义。

10.2.2 对数据结构的分析

本资料有两个实验因素——回心草的不同单体成分和给药剂量，全部实验条件由两因素各水平全面组合而成，每个实验条件下仅获得家兔的平均冠脉流量。就目前的资料而言，无法获得各实验条件下独立重复实验的数据，因此该资料应为双因素无重复实验设计定量资料。若表 10-2 中给出的数据不是各实验条件下的平均值，而是每只家兔的原始冠脉流量数据，则此资料应为“两因素析因设计定量资料”或“两因素嵌套设计定量资料”。

表 10-2 回心草各成分对缺血缺氧后兔离体心脏血流动力学的影响

| 单体成分 | 冠脉流量(\bar{x} , ml) | | | |
|-------|-----------------------|-------|------|------|
| | 剂量: | 高剂量 | 中剂量 | 低剂量 |
| 胡椒碱 | | 8.40 | 8.47 | 7.73 |
| 胡椒酸甲酯 | | 9.93 | 7.53 | 8.03 |
| 咖啡酸甲酯 | | 10.73 | 6.47 | 9.47 |

10.2.3 分析目的与统计分析方法的选择

研究者欲分析回心草的不同单体成分及给药剂量对冠脉流量的影响之间的差别是否有统计学意义。对此资料的统计分析方法与随机区组设计定量资料的分析方法相同，即资料满足参数检验的前提条件时，可采用随机区组设计定量资料方差分析；资料不满足参数检验的前提条件时，可采用 Friedman 秩和检验。

10.2.4 SAS 程序

SAS 程序名为 sastjfx10_2.sas。

```
data sastjfx10_2; /* 1 */
  do a=1 to 3;
    do b=1 to 3;
      input y @@ ;
      output;
    end;
  end;
  cards;
    8.40 8.47 7.73
    9.93 7.53 8.03
    10.73 6.47 9.47
  ;
run;
proc univariate normal noprint;
data = sastjfx10_2
  var y; /* 2 */
  class a;
  output out = set1 normal
         =w probn =p;
run;
proc univariate normal noprint
data = sastjfx10_2; /* 3 */
  var y;
  class b;
  output out = set2 normal
         =w probn =p;
run;
ods html;
data c; /* 4 */

proc glm data = sastjfx10_2; /* 5 */
  class a;
  model y = a/ss3;
  means a/hovtest;
  ods output HOVFTest = set3;
run;
proc glm data = sastjfx10_2; /* 6 */
  class b;
  model y = b/ss3;
  means b/hovtest;
  ods output HOVFTest = set4;
run;
data d; /* 7 */
  set set3 set4;
  if Source = 'Error' then delete;
  file print ods = (variables =
(Source FValue ProbF));
  title 'This is the results of Homo-
geneity of Variance test';
  put_ods_;
run;
title;
proc glm data = sastjfx10_2; /* 8 */
  class a b;
  model y = a b/ss3;
  means b/snk;
run;
ods html close;
```



```

set set1 set2;
file print ods = (variables
                  = (a b w p));
title 'This is the results of
      Normality test';
put_ods_;
run;
title;

```

程序中 a 代表“单体成分”，b 代表“剂量”，y 代表观测指标“冠脉流量”。其他与前一节中随机区组设计定量资料方差分析一致，此处不再解释。

10.2.5 主要分析结果及解释

| This is the results of Normality test | | | |
|---------------------------------------|---|---------------------------------|-----------------------------------|
| a | b | the normality test statistic, y | p-value of normality test stat, y |
| 1 | . | 0.8202516477 | 0.1638226536 |
| 2 | . | 0.8981288981 | 0.3795703266 |
| 3 | . | 0.9473189677 | 0.5576993591 |
| . | 1 | 0.9683167257 | 0.6582277205 |
| . | 2 | 0.9988014383 | 0.9338669773 |
| . | 3 | 0.8748266297 | 0.3093391845 |

这是分别对实验因素 a 和 b 所做的正态性检验结果，第一列为实验因素 a 及其 3 个水平，第二列为实验因素 b 及其 3 个水平，第三列为正态性检验的统计量 W 值，第四列为统计量 W 对应的 P 值。由第四列可以看出，a、b 两因素各水平对应的正态性检验结果 P 值均大于 0.05，可认为资料满足正态性要求。

| This is the results of Homogeneity of Variance test | | |
|---|---------|--------|
| Source | F Value | Pr > F |
| a | 2.63 | 0.1513 |
| b | 0.25 | 0.7857 |

这是对 a、b 两因素各水平进行方差齐性检验的结果，对应的 P 值分别为 0.1513、0.7857，均大于 0.05，可认为满足方差齐性条件。

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|--------|----|-------------|-------------|---------|--------|
| a | 2 | 0.71882222 | 0.35941111 | 0.25 | 0.7921 |
| b | 2 | 7.30162222 | 3.65081111 | 2.51 | 0.1966 |

这是双因素无重复实验设计一元定量资料方差分析的结果，a、b 两因素对应的 P 值分别为 0.7921、0.1966，说明两因素各水平对观测指标的影响之间的差异都没有统计学意义。即回心草的不同单体成分及给药剂量对冠脉流量的影响是相同的。

SAS 程序运行时，还分别给出了 a 因素、b 因素各水平两两比较的结果，由于两因素各水平对观测指标的影响之间的差异都没有统计学意义，所以两两比较的结果就无须查看了。

10.3 平衡不完全随机区组设计一元定量资料方差分析

10.3.1 问题与数据

【例 10-3】 某研究者欲研究服用 6 种不同剂量的维生素 D 对治疗佝偻病的效果差异是否有统计学意义, 选择了 15 窝白鼠, 每窝 4 只, 全部建造佝偻病模型。由于每窝中的白鼠数量少于维生素 D 的剂量水平数, 研究者采用了某种实验设计使接受各种剂量的维生素 D 处理的小鼠数相同, 并采用满分 20 分制对疗效进行评分, 结果见表 10-3。请进行合适的统计分析。

表 10-3 6 种不同剂量的维生素 D 治疗佝偻病的疗效评分

| 窝别 | 疗效评分(分) | | | | | | |
|-----|-----------|-----|-----|-----|-----|-----|-----|
| | 维生素 D 剂量: | I | II | III | IV | V | VI |
| 1 | | 6 | | 9 | 3 | 8 | |
| 2 | | 4 | | 13 | | 5 | 12 |
| ... | | ... | ... | ... | ... | ... | ... |
| 15 | | 11 | 9 | | | 3 | 15 |

10.3.2 对数据结构的分析

本资料中, 维生素 D 剂量的水平数 $a=6$, 小鼠窝别数 $b=15$, 每一窝别中的小鼠接受的处理数 $k=4$, 接受任一种剂量的维生素 D 处理的小鼠数 $r=10$ 。可以计算出本资料满足 $ar=bk$, 同时满足 $\lambda = r(k-1)/(a-1) = 10 \times (4-1)/(6-1) = 6$ 为整数, 故本资料应为平衡不完全区组设计一元定量资料。

10.3.3 分析目的与统计分析方法的选择

本资料为平衡不完全区组设计一元定量资料, 分析时应采用其相应的方差分析进行处理。

10.3.4 SAS 程序

设资料满足参数检验的前提条件, 现进行平衡不完全区组设计一元定量资料方差分析, SAS 程序名为 sastjfx10_3A.sas。

```
data sastjfx10_3A; /* 1 */
do litter=1 to 15;
do dose=1 to 6;
input y @@ ;
output;
end; end;
cards;
6 . 9 3 8 .
4 . 13 . 5 12
.....
11 9 . . 3 15
;
run;
```

```
ods html;
proc glm; /* 2 */
class litter dose;
model y=litter dose/ss3;
means dose/snk;
run;
ods html close;
```

程序第 1 步建立数据集, litter 代表区组因素“窝别”, dose 代表“维生素 D 剂量”, y 代表

观测指标“疗效评分”，“.”表示缺失值。第 3 个数据行以一串省略号表示中间未录入的数据，目的是节省篇幅，本章后面相似的程序中数据步也采用此方式录入数据。第 2 步调用 glm 过程进行方差分析，并对维生素 D 剂量各水平治疗佝偻病的疗效评分进行两两比较。

假设例 10-3 资料不满足参数检验的前提条件，现进行 Durbin 检验，SAS 程序见 sastjfx10_3B.sas。

```

data sastjfx10_3B; /* 1 */
  do litter = 1 to 15;
    do dose = 1 to 6;
      input y @@ ;
      output;
    end; end;
  cards;
6   .   9   3   8   .
4   .  13   .   5  12
.....
11  9   .   .   3  15
;
run;

ods html;
proc freq; /* 2 */
  tables litter* dose* y/noprint
  cmh2 scores = rank;
run;
ods html close;

```

10.3.5 主要分析结果及解释

下面是程序 sastjfx10_3A.sas 的输出结果。

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|-----------------|----|----------------|-------------|---------|--------|
| Model | 19 | 417.4472222 | 21.9709064 | 2.26 | 0.0148 |
| Error | 40 | 388.4861111 | 9.7121528 | | |
| Corrected Total | 59 | 805.9333333 | | | |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|--------|----|-------------|-------------|---------|--------|
| litter | 14 | 275.3138889 | 19.6652778 | 2.02 | 0.0409 |
| dose | 5 | 147.0138889 | 29.4027778 | 3.03 | 0.0207 |

这是输出结果的第一部分。可以看出：窝别和维生素 D 剂量的不同水平对疗效的影响之间的差别均有统计学意义，即不同窝别、不同维生素 D 剂量对疗效的影响是不完全相同的。另外，可以看出模型项对应的离均差平方和(417.4472222)不等于窝别(275.3138889)及维生素 D 剂量(147.0138889)的离均差平方和之合计，因为这里给出的是窝别和维生素 D 剂量调整后的离均差平方和。

Student-Newman-Keuls Test for y

Means with the same letter
are not significantly different.

| SNK Grouping | Mean | N | dose |
|--------------|--------|----|------|
| A | 10.700 | 10 | 6 |
| B | 9.800 | 10 | 3 |
| B | 8.600 | 10 | 5 |
| B | 7.600 | 10 | 2 |
| B | 7.400 | 10 | 1 |

B 6.100 10 4

这是输出结果的第二部分，是不同剂量水平对疗效影响的差异进行两两比较的结果。由“SNK Grouping”列可知，维生素 D 剂量 6 与剂量 4 对疗效影响之间的差异有统计学意义，其他任意两种剂量对疗效影响之间的差异均无统计学意义。由两种剂量疗效的均数对比(10.700 > 6.100)可知，维生素 D 剂量 6 的疗效优于剂量 4。

下面是程序 `sastjfx10_3B.sas` 的输出结果。

| The FREQ Procedure | | | | |
|---|------------------------|----|---------|--------|
| Summary Statistics fordose by y | | | | |
| Controlling for litter | | | | |
| Cochran-Mantel-Haenszel Statistics(Based on Rank Scores) | | | | |
| Statistic | Alternative Hypothesis | DF | Value | Prob |
| 1 | Nonzero Correlation | 1 | 7.6986 | 0.0055 |
| 2 | Row Mean Scores Differ | 5 | 15.0273 | 0.0102 |

查看上表中的行平均得分差异统计量，其值为 15.0273，对应的 P 值为 0.0102，说明不同剂量的维生素 D 对治疗佝偻病的效果差异是有统计学意义的，即这 6 种剂量的维生素 D 治疗佝偻病的效果不相同或不完全相同。

10.4 拉丁方设计一元定量资料方差分析

10.4.1 问题与数据

【例 10-4】 某研究者欲比较不同浓度的氯化钠溶液(1%、2%、4%、8%、16%，分别以数字 1、2、3、4、5 代替)静脉注射对家兔血压的影响，选取 5 只体重 2~3kg 的家兔进行实验，每只家兔分别注射不同浓度的氯化钠溶液各 1 次，注射量以 2mL/kg 计算，每次注射 10 秒(s)，观察并记录每次注射前后血压的最大变化值。两次注射之间间隔一段时间，使家兔血压恢复至初始水平，实验结果见表 10-4。已知因素间交互作用可以忽略不计，请判定资料设计类型，并进行合适的统计分析。

表 10-4 家兔静脉注射 5 种不同浓度 NaCl 溶液后血压升高值

| 兔 号 | NaCl 溶液浓度及血压升高值 (kPa) | | | | | |
|-----|-----------------------|---------|---------|---------|---------|---------|
| | 注射次序: | 1 | 2 | 3 | 4 | 5 |
| 1 | | 3(1.26) | 1(0.58) | 5(4.23) | 4(2.30) | 2(0.76) |
| 2 | | 2(0.73) | 5(4.15) | 4(2.19) | 3(1.56) | 1(0.18) |
| 3 | | 1(0.31) | 4(2.53) | 3(1.32) | 2(0.81) | 5(4.53) |
| 4 | | 4(2.01) | 2(0.47) | 1(0.51) | 5(4.33) | 3(1.40) |
| 5 | | 5(4.13) | 3(1.26) | 2(0.70) | 1(0.60) | 4(2.30) |

10.4.2 对数据结构的分析

本资料中，受试对象为 5 只家兔，实验因素为 NaCl 溶液浓度，行区组因素为兔号，列区组因素为注射次序，观测指标为血压升高值，故此资料为拉丁方设计一元定量资料。

10.4.3 分析目的与统计分析方法的选择

此资料为拉丁方设计一元定量资料，应选用其相应的方差分析进行处理。

10.4.4 SAS 程序

SAS 程序名为 sastjfx10_4.sas。

```

data sastjfx10_4;          /* 1 */
  do rabbit = 1 to 5;
  do order = 1 to 5;
  input dose y @@ ;
  output; end; end;
cards;
3 1.26 1 0.58 5 4.23 4 2.30 2 0.76
2 0.73 5 4.15 4 2.19 3 1.56 1 0.18
1 0.31 4 2.53 3 1.32 2 0.81 5 4.53
4 2.01 2 0.47 1 0.51 5 4.33 3 1.40
5 4.13 3 1.26 2 0.70 1 0.60 4 2.30
;
run;

ods html;
proc anova;                /* 2 */
  class rabbit order dose;
  model y = rabbit order dose;
  means dose / snk;
run;
ods html close;

```

程序第 1 步建立数据集，rabbit 代表“家兔编号”，order 代表“注射次序”，dose 代表“不同浓度的氯化钠溶液”，y 代表“血压升高值”。第 2 步调用 anova 过程进行方差分析，并对 5 种浓度的氯化钠溶液的升血压效果进行两两比较。

10.4.5 主要分析结果及解释

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|--------|----|-------------|-------------|---------|--------|
| rabbit | 4 | 0.07540000 | 0.01885000 | 0.75 | 0.5767 |
| order | 4 | 0.14012000 | 0.03503000 | 1.39 | 0.2941 |
| dose | 4 | 48.07492000 | 12.01873000 | 478.26 | <.0001 |

这是输出结果的第一部分。可以看出，不同家兔和不同注射顺序对血压升高值影响之间的差异均无统计学意义，而不同浓度的氯化钠溶液对血压升高值影响之间的差异则有统计学意义 ($F = 478.26$, $P < 0.0001$)。

Student-Newman-Keuls Test for y

Means with the same letter
are not significantly different.

| SNK Grouping | Mean | N | drug |
|--------------|--------|---|------|
| A | 4.2740 | 5 | 5 |
| B | 2.2660 | 5 | 4 |
| C | 1.3600 | 5 | 3 |
| D | 0.6940 | 5 | 2 |
| E | 0.4360 | 5 | 1 |

这是结果输出的第二部分。由第一部分结果可知，不同浓度的氯化钠溶液对血压升高值的影响不全相同，这里给出了 5 种浓度的氯化钠溶液对血压升高值影响两两比较的结果。由“SNK Grouping”列可以看出，5 种浓度的氯化钠溶液对血压升高值影响之间的差异均有统计学

意义。由均值大小可知各浓度氯化钠溶液对血压升高值影响的程度由大到小分别为 $16\% > 8\% > 4\% > 2\% > 1\%$ 。

其实,从第一部分方差分析的结果可以看出,不同家兔和不同注射顺序对血压升高的影响之间的差异没有统计学意义,可在 SAS 程序中的 class 和 model 语句中删除 rabbit 和 order,这样可以增大误差项的自由度,使结果更加稳定可靠。重新运行程序,得检验结果如下:

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|--------|----|-------------|-------------|---------|--------|
| dose | 4 | 48.07492000 | 12.01873000 | 464.87 | <.0001 |

Means with the same letter
are not significantly different.

| SNK Grouping | Mean | N | drug |
|--------------|--------|---|------|
| A | 4.2740 | 5 | 5 |
| B | 2.2660 | 5 | 4 |
| C | 1.3600 | 5 | 3 |
| D | 0.6940 | 5 | 2 |
| E | 0.4360 | 5 | 1 |

结果与前面相同,专业结论同上。

10.5 二阶段交叉设计一元定量资料方差分析

10.5.1 问题与数据

【例 10-5】某公司原生产药物 B,现对其剂型进行改造,生产出药物 A,欲评价 A 药与 B 药是否具有相同的生物利用度,以 AUC(血药浓度-时间曲线下面积)作为评价指标,选取 24 名受试者并将他们随机均分成两组,用药顺序为:第一组受试者在第一、第二实验周期分别接受 A 药与 B 药,以 AB 表示;第二组受试者在第一、第二实验周期分别接受 B 药与 A 药,以 BA 表示。在两个实验周期中设立了“洗脱期”,实验结果见表 10-5。请判断资料的类型并进行合适的统计分析。

表 10-5 A、B 两种剂型药物影响下的血药浓度-时间曲线下面积

| 用药顺序 | 受试者号 | AUC((μg/mL) · h) | | 用药顺序 | 受试者号 | AUC((μg/mL) · h) | |
|------|------|--------------------|------|------|------|--------------------|-------|
| | | 周期 1 | 周期 2 | | | 周期 1 | 周期 2 |
| AB | 1 | 84.4 | 70.9 | BA | 3 | 66.6 | 63.6 |
| | 2 | 101.0 | 92.7 | | 5 | 117.0 | 107.3 |
| | ... | ... | ... | | ... | ... | ... |
| | 23 | 61.9 | 51.9 | | 24 | 74.7 | 76.0 |

10.5.2 对数据结构的分析

本资料中有一个实验因素“药物剂型”,它有两个水平(即 A 剂型和 B 剂型)。24 名受试者被随机均分成两组,以 AB、BA 两种相反的顺序接受两种剂型药物的处理,实验周期中设有洗脱期。所以,此资料应为二阶段交叉设计一元定量资料。

10.5.3 分析目的与统计分析方法的选择

因本资料为二阶段交叉设计一元定量资料,故应采用相应的方差分析来处理。

10.5.4 SAS 程序

SAS 程序名为 sastjfx10_5.sas。

```

data sastjfx10_5;          /* 1 */      ;
  do subject =1 to 24;
    do order =1 to 2;
      input drug $ y@@ ;
      output;
    end; end;
  cards;
A 84.4 B 70.9 A 101.0
B 92.7 A 72.7 B 99.0
. . . . .
A 61.5 B 74.7 A 76.0
                                /* 2 */
run;
ods html;
proc anova;
class order subject drug;
model y=order subject drug;
run;
ods html close;

```

程序第 1 步建立数据集, subject 代表“受试者”, order 代表“用药次序”, drug 代表“药物种类”, y 为观测指标“AUC(($\mu\text{g/ml}$) $\cdot\text{h}$)”。其他部分与拉丁方设计定量资料方差分析相似, 此处不再赘述。另外, 在 SAS 9.2 中, 已经可以用 ttest 过程做二阶段交叉设计一元定量资料方差分析了。其基本语法为:

```

proc ttest;
  var outcome1 outcome2/crossover = (treat1 treat2);
run;

```

其中, outcome1 和 outcome2 分别代表第一阶段和第二阶段受试对象的观测值, treat1 和 treat2 分别代表第一阶段和第二阶段受试对象接受的处理。当然, 在录入原始数据时, 其形式与程序 sastjfx10_5.sas 会有所差异。如本例, 可直接将 do 循环中语句换成“input period1 \$ y1 period2 \$ y2@@;”。调用 ttest 过程进行统计分析时, 将上述语法修改为“var y1 y2/ crossover = (period1 period2);”即可。综合来说, 调用 ttest 过程做二阶段交叉设计一元定量资料方差分析要比 anova 简单一些, 录入数据更直观, 有兴趣的读者可以尝试一下。

10.5.5 主要分析结果及解释

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|---------|----|-------------|-------------|---------|--------|
| order | 1 | 82.68750 | 82.68750 | 0.99 | 0.3298 |
| subject | 23 | 14190.92000 | 616.99652 | 7.41 | <.0001 |
| drug | 1 | 0.24083 | 0.24083 | 0.00 | 0.9576 |

这是输出的结果。可以看出, 不同实验顺序和药物间 AUC 的差别均无统计学意义, 而受试者个体的差异则存在统计学意义($F=7.41$, $P<0.001$), 说明 A 药和 B 药生物利用度相等。

10.6 析因设计一元定量资料方差分析

10.6.1 问题与数据

【例 10-6】 某研究者欲研究 IL-11 药对 5.5Gy 照射小鼠骨髓造血细胞周期(G_0/G_1 期)的影响, 选取 45 只小鼠并将其完全随机地均分成 3 组, 每组 15 只, 分别在 5.5Gy 剂量照射前

给 IL-11 药、照射后给 IL-11 药和照射对照(即不给 IL-11 药)。每一组中的 15 只小鼠随机等分成 3 组,分别在照后 6h、12h、24h 三个时间点上处死,测量其骨髓造血细胞周期(G_0/G_1 期)。实验结果见表 10-6, 请进行相应的统计分析。

表 10-6 IL-11 药物对 5.5Gy 照射小鼠骨髓造血细胞周期(G_0/G_1 期)的影响

| 给药时机 | 骨髓造血细胞周期(G_0/G_1 期)(单位) | | | |
|-------|----------------------------|-------|-------|-------|
| | 处死时间(h): | 6 | 12 | 24 |
| 照射前给药 | | 65.80 | 86.83 | 95.88 |
| | | ... | ... | ... |
| | | 63.69 | 86.10 | 90.84 |
| 照射后给药 | | 59.30 | 80.69 | 95.53 |
| | | ... | ... | ... |
| | | 60.46 | 83.30 | 93.67 |
| 照射对照 | | 65.58 | 86.56 | 95.23 |
| | | ... | ... | ... |
| | | 64.89 | 85.47 | 95.65 |

10.6.2 对数据结构的分析

实验中涉及两个实验因素：“给药时机”和“处死时间”，受试对象为小鼠。表面上看，此设计像具有一个重复测量的两因素设计，这样看是把“处死时间”当成了重复测量因素。其实，不同时间点上获得的数据没有关联，不是来自于同一只小鼠，因为每只小鼠身上只能测到一个数据。所以，此资料不是具有一个重复测量的两因素设计定量资料。由于实验条件为两因素各水平的全面组合，研究者在每种实验条件下均进行了 5 次独立重复实验，在没有专业依据认为这两个实验因素对观测结果的影响重要程度不同的情况下，此资料应为两因素析因设计定量资料。

10.6.3 分析目的与统计分析方法的选择

对本资料，应采用两因素析因设计定量资料方差分析来处理。这样既可分析两实验因素的主效应，也可分析二者之间的交互效应。

10.6.4 SAS 程序

SAS 程序名为 sastjfx10_6.sas。

```
data sastjfx10_6;          /* 1 */
  do treat = 1 to 3;
  do rep = 1 to 5;
  do time = 1 to 3;
  input outcome @@ ;
  output;
  end; end; end;
  cards;
65.80 86.83 95.88 66.72 84.69 93.67
66.59 84.69 96.67 62.82 87.92 94.93
. . . . .
64.89 85.47 95.65
;
run;
```

```
ods html;
proc glm;                  /* 2 */
  class treat time;
  model outcome = treat | time;
  lsmeans treat * time / tdiff pdiff;
run;
ods html close;
```


程序第 1 步建立数据集, treat 代表“给药时机”, rep 代表“受试对象个体”, time 代表“处死时间”, outcome 代表“骨髓造血细胞周期(G_0/G_1 期)(单位)”。第 2 步进行两因素析因设计定量资料方差分析, 其中“treat|time”等价于“treat time treat * time”, TDIFF 和 PDIFF 则分别用来输出两两比较的 t 值和相应的 P 值。

10.6.5 主要分析结果及解释

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|--------------|----|-------------|-------------|---------|--------|
| treat | 2 | 109.569391 | 54.784696 | 14.19 | <.0001 |
| time | 2 | 7828.731858 | 3914.365929 | 1013.99 | <.0001 |
| treat * time | 4 | 41.598996 | 10.399749 | 2.69 | 0.0462 |

这是输出结果的第一部分。由方差分析结果可以看出, 两个实验因素“给药时机”和“处死时间”及其交互作用均有统计学意义, 即不同给药时机、不同处死时间上骨髓造血细胞周期的值是不完全相同的。

Least Squares Means

| treat | time | outcome LSMEAN | LSMEAN Number |
|-------|------|----------------|---------------|
| 1 | 1 | 65.1240000 | 1 |
| 1 | 2 | 86.0460000 | 2 |
| 1 | 3 | 94.3980000 | 3 |
| 2 | 1 | 59.1140000 | 4 |
| 2 | 2 | 83.0220000 | 5 |
| 2 | 3 | 93.7300000 | 6 |
| 3 | 1 | 64.8900000 | 7 |
| 3 | 2 | 85.4720000 | 8 |
| 3 | 3 | 95.6480000 | 9 |

Least Squares Means for Effect treat * time

t for H_0 : $LSMean(i) = LSMean(j) / Pr > |t|$

Dependent Variable: outcome

| i/j | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 1 | | -16.84 | -23.56 | 4.84 | -14.40 | -23.02 | 0.19 | -16.37 | -24.56 |
| | | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | 0.8517 | <.0001 | <.0001 |
| 2 | 16.84 | | -6.72 | 21.67 | 2.43 | -6.18 | 17.03 | 0.46 | -7.73 |
| | <.0001 | | <.0001 | <.0001 | 0.0200 | <.0001 | <.0001 | 0.6469 | <.0001 |
| 3 | 23.56 | 6.72 | | 28.39 | 9.15 | 0.54 | 23.75 | 7.18 | -1.01 |
| | <.0001 | <.0001 | | <.0001 | <.0001 | 0.5942 | <.0001 | <.0001 | 0.3212 |
| 4 | -4.84 | -21.67 | -28.39 | | -19.24 | -27.86 | -4.65 | -21.21 | -29.40 |
| | <.0001 | <.0001 | <.0001 | | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 |
| 5 | 14.40 | -2.43 | -9.15 | 19.24 | | -8.62 | 14.59 | -1.97 | -10.16 |
| | <.0001 | 0.0200 | <.0001 | <.0001 | | <.0001 | <.0001 | 0.0564 | <.0001 |
| 6 | 23.02 | 6.18 | -0.54 | 27.86 | 8.62 | | 23.21 | 6.65 | -1.54 |
| | <.0001 | <.0001 | 0.5942 | <.0001 | <.0001 | | <.0001 | <.0001 | 0.1315 |

| | | | | | | | | | |
|---|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 7 | -0.19 | -17.03 | -23.75 | 4.65 | -14.59 | -23.21 | | -16.56 | -24.75 |
| | 0.8517 | <.0001 | <.0001 | <.0001 | <.0001 | <.0001 | | <.0001 | <.0001 |
| 8 | 16.37 | -0.46 | -7.18 | 21.21 | 1.97 | -6.65 | 16.56 | | -8.19 |
| | <.0001 | 0.6469 | <.0001 | <.0001 | 0.0564 | <.0001 | <.0001 | | <.0001 |
| 9 | 24.56 | 7.73 | 1.01 | 29.40 | 10.16 | 1.54 | 24.75 | 8.19 | |
| | <.0001 | <.0001 | 0.3212 | <.0001 | <.0001 | 0.1315 | <.0001 | <.0001 | |

Note: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

这是输出结果的第二部分。本部分首先给出了各实验组在两两比较中的编号，见“LSMEAN Number”列。随后给出了各实验组两两比较的结果，行与列的交叉处即是这两组比较的结果，上行为 t 值，下行为对应的 P 值，形式上看两两比较用的是 t 检验，实际上是方差分析计算出来的。另由于篇幅排版限制，对输出结果进行了四舍五入，所以本部分结果比 SAS 实际输出结果保留的小数点后位数要少一些。由两两比较结果可知：除实验组 1 与实验组 7、实验组 2 与实验组 8、实验组 3 与实验组 6、实验组 3 与实验组 9、实验组 5 与实验组 8、实验组 6 与实验组 9 之间观测指标的差异无统计学意义外，其他各组间观测指标的差异均有统计学意义。

10.7 含区组因素的析因设计一元定量资料方差分析

10.7.1 问题与数据

【例 10-7】 为研究克拉霉素的抑菌效果，将 28 个短小芽孢杆菌平板依据菌株的来源不同分成了 7 个区组，每组 4 个平板，用随机的方式分配给标准药物高剂量组 (SH)、标准药物低剂量组 (SL)、克拉霉素高剂量组 (TH)、克拉霉素低剂量组 (TL)。给予不同的处理后，观测抑菌圈的直径，结果见表 10-7。问：这是什么设计类型？应如何进行统计分析？

表 10-7 28 个平板给予不同处理后的抑菌圈直径

| 区组 | 抑菌圈直径 (mm) | | | |
|-----|------------|-------|-------|-------|
| | SL | SH | TL | TH |
| 1 | 18.02 | 19.41 | 18.00 | 19.46 |
| 2 | 18.12 | 20.20 | 18.91 | 20.38 |
| ... | ... | ... | ... | ... |
| 7 | 18.23 | 19.94 | 18.06 | 19.54 |

10.7.2 对数据结构的分析

研究者在分配受试对象时，没有采用完全随机分配的方法，而是根据菌株的来源不同分成了 7 个区组，每个区组内的 4 个平板再随机分配给“药物种类”和“剂量大小”两个实验因素全面组合而成的 4 个实验组，且两个实验因素各水平组合而成的每一个实验条件下都进行了 7 次独立重复实验，因此该资料应为含区组因素的析因设计。

10.7.3 分析目的与统计分析方法的选择

分析时，若资料满足方差分析的前提条件，应采用含区组因素析因设计定量资料方差分析，将区组因素对实验结果的影响分离出来，从而更好地评价各实验因素及其交互作用的效应大小。

10.7.4 SAS 程序

SAS 程序名为 sastjfx10_7.sas。

```

data sastjfx10_7;          /* 1 */
  do block=1 to 7;
    do drug=1 to 2;
      do dose=1 to 2;
        input diameter@@ ;
        output;
      end; end; end;
  cards;
  18.02 19.41  18.00 19.46  18.12
  20.20 18.91  20.38 18.09  19.56
  18.21 19.64  18.30 19.41  18.24
  19.50 18.26  19.59 18.11  19.56
  18.02 20.12  18.13 19.60  18.23
  19.94 18.06  19.54
  ;
run;

ods html;
proc glm;                  /* 2 */
  class block drug dose;
  model diameter = block drug |
                        dose/SS3;
  lsmeans drug* dose/tdiff pdiff;
run;
ods html close;

```

程序第 1 步建立数据集, block 代表区组因素“菌株来源”, drug 代表“药物种类”, dose 代表“剂量”, diameter 代表观测指标“抑菌圈直径”。其他程序与析因设计定量资料方差分析 SAS 程序相似, 仅在 class 和 model 语句中多了一个区组项“block”。

10.7.5 主要分析结果及解释

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|-------------|----|-------------|-------------|---------|--------|
| block | 6 | 1.10022143 | 0.18337024 | 3.99 | 0.0103 |
| drug | 1 | 0.00017500 | 0.00017500 | 0.00 | 0.9515 |
| dose | 1 | 16.06657500 | 16.06657500 | 349.26 | <.0001 |
| drug * dose | 1 | 0.04888929 | 0.04888929 | 1.06 | 0.3162 |

这是输出结果的第一部分, 可以看出区组因素和给药剂量因素所对应的检验统计量和 P 值分别为 $F = 3.99$ 、 $P = 0.0103$, $F = 349.26$ 、 $P < 0.0001$, 即不同菌株来源和不同给药剂量对观测指标的影响之间的差异具有统计学意义。

Least Squares Means

| drug | dose | diameter LSMEAN | LSMEAN Number |
|------|------|-----------------|---------------|
| 1 | 1 | 18.1485714 | 1 |
| 1 | 2 | 19.7471429 | 2 |
| 2 | 1 | 18.2371429 | 3 |
| 2 | 2 | 19.6685714 | 4 |

Least Squares Means for Effect drug * dose

t for $H_0: \text{LSMean}(i) = \text{LSMean}(j) / \text{Pr} > |t|$

Dependent Variable: diameter

| i/j | 1 | 2 | 3 | 4 |
|-----|---|----------|----------|----------|
| 1 | | -13.9437 | -0.77257 | -13.2583 |
| | | <.0001 | 0.4498 | <.0001 |

| | | | | |
|---|----------|----------|----------|----------|
| 2 | 13.94369 | | 13.17112 | 0.685347 |
| | <.0001 | | <.0001 | 0.5019 |
| 3 | 0.772573 | -13.1711 | | -12.4858 |
| | 0.4498 | <.0001 | | <.0001 |
| 4 | 13.25835 | -0.68535 | 12.48577 | |
| | <.0001 | 0.5019 | <.0001 | |

Note: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

这是输出结果的第二部分，是全部实验组之间两两比较的结果。本部分首先给出了各实验组在两两比较中的编号，见“LSMEAN Number”列。随后给出了各实验组两两比较的结果，可以看出：除实验组 1 与实验组 3、实验组 2 与实验组 4 之间的差异无统计学意义外，其他各组间的比较均有统计学差异。综上所述，在控制药物剂量在低剂量或高剂量时，标准药物与克拉霉素对抑菌圈的直径的影响效果没有差异。

10.8 嵌套设计一元定量资料方差分析

10.8.1 问题与数据

【例 10-8】 某研究者进行绵马贯众及单芽狗脊贯众饮片的凝血时间对比实验，取昆明种小鼠 48 只，随机分成 8 组，分别为绵马贯众生品高低剂量组、炭品高低剂量组，单芽狗脊贯众生品高低剂量组、炭品高低剂量组。各组小鼠连续灌胃给药 3d，第 3d 给药 1h 后以毛细管法测定小鼠凝血时间(s)，数据见表 10-8。假设专业上认为对观测指标的影响，生品或炭品选择影响最大，然后是剂量的选择(高剂量或低剂量)，最后是药物的选择(绵马贯众或单芽狗脊贯众)。实验结果如下，请进行合适的统计分析。

表 10-8 药物对凝血时间的影响

| 炮制类型 | 剂量(g 生药 /kg 体重) | 凝血时间(s) | | | |
|------|--------------------|---------|--------|--------|--------|
| | | 药材种类: | 绵马贯众 | 单芽狗脊贯众 | |
| 生品 | 2.25 | | 155.59 | 182.04 | 185.11 |
| | | | 129.76 | 174.66 | 158.70 |
| | | | 167.10 | 139.73 | 159.24 |
| | 0.75 | | 154.71 | 185.22 | 137.97 |
| | | | 167.94 | 137.83 | 173.18 |
| | | | 172.75 | 173.65 | 164.19 |
| 炭品 | 2.25 | | 144.07 | 128.19 | 98.94 |
| | | | 150.35 | 91.98 | 108.09 |
| | | | 91.56 | 135.76 | 133.52 |
| | 0.75 | | 145.45 | 143.09 | 140.96 |
| | | | 152.37 | 122.15 | 122.08 |
| | | | 118.37 | 114.77 | 143.65 |

10.8.2 对数据结构的分析

研究者做了 8 组实验，乍一看，很像单因素 8 水平设计定量资料，实则不然。这 8 个实验组实际上涉及了三个实验因素：“药材种类”、“剂量”和“炮制类型”，且 8 个实验条件正是这

三个实验因素各水平的全面组合。由于有专业依据认为三个实验因素对观测指标的影响重要性不同, 炮制类型 > 剂量 > 药材种类, 那么此资料应为三因素嵌套设计一元定量资料。

10.8.3 分析目的与统计分析方法的选择

应选用与其设计类型相对应的方差分析处理此资料。

10.8.4 SAS 程序

SAS 程序名为 sastjfx10_8.sas。

```

data sastjfx10_8;          /* 1 */
  do drug = 1 to 2;
  do type = 1 to 2;
  do dose = 1 to 2;
  do rep = 1 to 6;
  input y @@ ;
  output;
  end; end; end; end;
  cards;
  155.59 182.04 129.76 174.66
  167.10 139.73 154.71 185.22
  . . . . .
  122.08 145.97 143.65 150.12
;
run;
proc sort;                  /* 2 */
  by type dose drug;
run;

ods html;
proc nested;                /* 3 */
  class type dose drug;
  var y;
run;
proc glm;                   /* 4 */
  class drug type dose;
  model y = type dose (type) drug
(dose type)/ssl;
  test h = type e = dose (type);
  test h = dose (type) e = drug (dose
type);
  means drug (dose type);
run;
ods html close;

```

程序第 1 步建立数据集, drug 代表“药材种类”, type 代表“炮制类型”, dose 代表“剂量”, rep 代表“受试对象个体”, y 代表观测指标“凝血时间”。第 2 步对数据集进行排序, 这是 nested 过程的要求。在调用 nested 过程之前, 需按照 nested 过程中 class 语句后因素的顺序对数据集进行排序, 即 sort 过程中 by 语句后的因素顺序必须与 nested 过程中 class 语句后因素的顺序保持一致。当然, 若数据集本身就是按照 nested 过程中 class 语句后因素的顺序建立的, 此过程可以省略。如本数据集以循环语句“do type = 1 to 2; do dose = 1 to 2; do drug = 1 to 2;”来录入数据集, 那么建立的数据集本身就是按照 type、dose、drug 的顺序排列的, 可以不必再调用 sort 过程。第 3 步调用 nested 过程进行方差分析, 此过程适用于分析平衡的嵌套设计定量资料(即各实验条件下进行相同次数独立重复实验的第一类嵌套设计, 各水平下重复实验次数相等且主因素各水平下嵌套的次因素水平数相等)。在 class 语句中, 各因素出现的顺序很重要, 要求按照因素重要程度由高到低排列, 即主因素在前, 次因素在后。第 4 步调用 glm 过程进行嵌套设计定量资料方差分析, 与 nested 过程不同的是, glm 适用范围更广, 它可以分析不平衡的嵌套设计定量资料, 也不需要事先调用 sort 过程, 当然 class 语句中各因素的排列顺序也可随意安排。但在 model 语句中需表明各因素的嵌套关系, 如本例中 type 为主因素, dose 为次因素, drug 为次次因素, 那么在 model 语句中就需要写成 y = type dose (type) drug (dose type); dose (type) 表示 dose 嵌套于 type 因素之下, drug (dose type) 表示 drug 嵌套于 dose 和 type 因素之下。此外, glm 过程还需要用 test 语句指明分析各因素主效应时所采用的误差项。

10.8.5 主要分析结果及解释

| The NESTED Procedure | | | | | | | | |
|---|----|----------------|---------|--------|--------------|-------------|--------------------|------------------|
| Nested Random Effects Analysis of Variance for Variable y | | | | | | | | |
| Variance Source | DF | Sum of Squares | F Value | Pr > F | Error Term | Mean Square | Variance Component | Percent of Total |
| Total | 47 | 31311 | | | | 666.18 | 1019.19 | 100.00 |
| type | 1 | 15127 | 19.87 | 0.0468 | dose | 15127 | 598.56 | 58.73 |
| dose | 2 | 1522.81 | 14.94 | 0.0139 | drug | 761.41 | 59.20 | 5.81 |
| drug | 4 | 203.92 | 0.14 | 0.9659 | Error | 50.98 | -51.74 | 0.00 |
| Error | 40 | 14457 | | | | 361.42 | 361.42 | 35.46 |
| y Mean | | | | | 145.92729167 | | | |
| Standard Error of y Mean | | | | | 17.75229167 | | | |

这是输出结果的第一部分，是 nested 过程产生的结果。可以看出：炮制类型和剂量各水平对凝血时间的影响之间的差异有统计学意义，对应的统计量值分别为 $F = 19.87$ 、 $F = 14.94$ ，相应的 P 值分别为 $P = 0.0468$ 、 $P = 0.0139$ 。在“Error Term”列可以查看计算各因素 F 值所用的误差项，如分析炮制类型时所用的误差项为剂量，分析剂量时所用的误差项为药物种类，分析药物种类时所用的误差项为模型的误差。由于篇幅限制，上述结果中对小数点后位数较多的数值进行了四舍五入。

| The GLM Procedure | | | | | |
|--------------------|----|-------------|-------------|---------|--------|
| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
| type | 1 | 15126.90525 | 15126.90525 | 41.85 | <.0001 |
| dose(type) | 2 | 1522.81212 | 761.40606 | 2.11 | 0.1349 |
| drug(type * dose) | 4 | 203.91646 | 50.97911 | 0.14 | 0.9659 |

这是输出结果的第二部分，是 glm 过程产生的结果，给出了模型拟合的有关信息和方差分析表。需要注意的是，此方差分析表仅有最后一行的信息是正确的，即对药物种类检验的结果，其中 $F = 0.14$ ， $P = 0.9659$ ，说明不同药物造成的凝血时间的差异没有统计学意义。

Tests of Hypotheses Using the Type I MS for dose(type) as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|--------|----|-------------|-------------|---------|--------|
| type | 1 | 15126.90525 | 15126.90525 | 19.87 | 0.0468 |

Tests of Hypotheses Using the Type I MS for drug(type * dose) as an Error Term

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|-------------|----|------------|-------------|---------|--------|
| dose(type) | 2 | 1522.81212 | 761.40606 | 14.94 | 0.0139 |

| Level of drug | Level of type | Level of dose | N | Mean | Std Dev |
|---------------|---------------|---------------|---|------------|------------|
| 1 | 1 | 1 | 6 | 158.146667 | 20.3728090 |
| 2 | 1 | 1 | 6 | 159.208333 | 17.2696780 |
| 1 | 1 | 2 | 6 | 165.350000 | 16.7076928 |
| 2 | 1 | 2 | 6 | 172.013333 | 19.6179007 |
| 1 | 2 | 1 | 6 | 123.651667 | 25.8065383 |

| | | | | | |
|---|---|---|---|------------|------------|
| 2 | 2 | 1 | 6 | 120.300000 | 19.3056303 |
| 1 | 2 | 2 | 6 | 132.700000 | 16.0969550 |
| 2 | 2 | 2 | 6 | 136.048333 | 14.7068140 |

这是输出结果的第三部分，即由 glm 过程中两个 test 语句和 means 语句得出的结果。前两个表分别给出了对炮制类型和剂量进行检验的结果，其统计量取值分别为 $F = 19.87$ 、 $F = 14.94$ ，对应的 P 值分别为 $P = 0.0468$ 、 $P = 0.0139$ 。说明不同炮制类型和不同剂量造成的凝血时间的差异有统计学意义，根据最后给出的均数表可以看出，生品的凝血时间一般大于炭品，低剂量药品的凝血时间一般大于高剂量药品，说明炭品比生品的凝血时间短，高剂量药品比低剂量药品凝血时间短。

请注意：若读者想计算炮制类型、剂量、药物种类各自水平对应的具体均数，在此 glm 过程中是无法实现的，因为 glm 过程要求 means 语句后的效应成分必须在之前的 model 语句中出现过。所以，需重新构建 model 语句，具体做法如下：

```
proc glm;
  class drug type dose;
  model y = type dose drug / ssl;
  means type dose drug;
run;
```

这样就可以输出三个实验因素各自水平的均数及标准差，当然此过程给出的统计分析是错误的，读者可不予关注。此过程计算的各因素各水平均数如下：

| Level of | N | Mean | Std Dev |
|----------|----|------------|------------|
| type | | | |
| 1 | 24 | 163.679583 | 18.2060576 |
| 2 | 24 | 128.175000 | 19.2919026 |
| Level of | N | Mean | Std Dev |
| dose | | | |
| 1 | 24 | 140.326667 | 27.0900639 |
| 2 | 24 | 151.527917 | 23.7064642 |
| Level of | N | Mean | Std Dev |
| drug | | | |
| 1 | 24 | 144.962083 | 25.7584061 |
| 2 | 24 | 146.892500 | 26.3797236 |

10.9 裂区设计一元定量资料方差分析

10.9.1 问题与数据

【例 10-9】某研究者欲研究蛇毒的抑瘤作用，选用 48 只小白鼠，根据重要非实验因素先将其分为 3 个区组，每个区组 16 只小白鼠，然后将每个区组中的 16 只小白鼠随机等分为 4 组，分别接种 4 种不同的瘤株。接种成功后，再随机决定每个区组中接种每种瘤株的 4 只小白鼠分别经腹腔注射 4 种浓度的蛇毒，观测相应的瘤重，实验结果见表 10-9。请判断此资料所取自的实验设计类型，并对资料做相应的统计分析。

表 10-9 不同瘤株和不同浓度的蛇毒对小鼠抑瘤效果的影响

| 瘤株 | 蛇毒浓度 | 瘤重 (g) | | | 瘤株 | 蛇毒浓度 | 瘤重 (g) | | | | |
|------|-------|--------|------|------|------|------|--------|-----|------|------|------|
| | | 区组: | 1 | 2 | | | 3 | 区组: | 1 | 2 | 3 |
| S180 | 0 | | 0.80 | 0.76 | 0.36 | EC | 0 | | 0.31 | 0.55 | 0.32 |
| | 0.030 | | 0.36 | 0.26 | 0.31 | | 0.030 | | 0.20 | 0.15 | 0.20 |
| | 0.050 | | 0.17 | 0.28 | 0.16 | | 0.050 | | 0.38 | 0.18 | 0.26 |
| | 0.075 | | 0.28 | 0.13 | 0.11 | | 0.075 | | 0.25 | 0.21 | 0.14 |
| HS | 0 | | 0.74 | 0.43 | 0.57 | ARS | 0 | | 0.48 | 0.57 | 0.33 |
| | 0.030 | | 0.50 | 0.46 | 0.32 | | 0.030 | | 0.18 | 0.30 | 0.29 |
| | 0.050 | | 0.42 | 0.20 | 0.20 | | 0.050 | | 0.44 | 0.27 | 0.27 |
| | 0.075 | | 0.36 | 0.26 | 0.32 | | 0.075 | | 0.22 | 0.30 | 0.37 |

10.9.2 对数据结构的分析

本实验中，实验因素“瘤株”先出现在实验中，将每个区组中的 16 个小鼠分成 4 个一级单位，每个一级单位中有 4 只小鼠，均接种其中的某一种瘤株。接种成功后，然后将每个区组中的每个一级单位再分成 4 个二级单位(对每一个区组来说，每个二级单位仅含 1 只小鼠)，再随机决定它们分别被注射 4 种不同浓度的蛇毒。也就是说，实验因素“瘤株”和“蛇毒浓度”进入实验有先后之分，同时实验中涉及一个区组因素，因此这是含一个区组因素的裂区设计。

10.9.3 分析目的与统计分析方法的选择

对本资料，应选用裂区设计定量资料方差分析。

10.9.4 SAS 程序

SAS 程序名为 sastjfx10_9.sas。

```
data sastjfx10_9;                                /* 1 */
do type=1 to 4;
do dose=1 to 4;
do block=1 to 3;
input y @@ ;
output;
end; end; end;
cards;
0.80 0.76 0.36 0.36 0.26 0.31
0.17 0.28 0.16 0.28 0.13 0.11
. . . . .
0.44 0.27 0.27 0.22 0.30 0.37
;
run;
ods html;
```

```
proc glm;                                        /* 2 */
class type dose block;
model y = type block type* block
dose type* dose/ss3;
test h = type block e = type* block;
lsmeans type* dose/tdiff pdiff;
run;
proc mixed;                                    /* 3 */
class type dose block;
model y = type dose type* dose;
random block block* type;
lsmeans type* dose/pdiff;
run;
ods html close;
```

程序第 1 步建立数据集，type 代表“瘤株”，dose 代表“蛇毒浓度”，block 代表“区组”，y 代表观测指标“瘤重”。第 2 步调用 glm 过程进行裂区设计定量资料(含一个区组因素)的方差分析，参照前述方差分析表，在 model 语句中列出所有需要分析的效应，并在随后的 test 语句中对分析 type 和 block 两主效应所用的误差项进行明确。第 3 步调用 mixed 过程进行裂区设计定量资料(含一个区组因素)的方差分析。

当资料属于平衡资料时, glm 与 mixed 过程的计算结果是相同的;当资料属于不平衡资料时, 二者的计算结果不完全相同;有时 glm 过程提示不可估计, 此时需调用 mixed 过程。

10.9.5 主要分析结果及解释

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|--------------|----|------------|-------------|---------|--------|
| type | 3 | 0.11087292 | 0.03695764 | 3.59 | 0.0283 |
| block | 2 | 0.07605000 | 0.03802500 | 3.69 | 0.0399 |
| type * block | 6 | 0.06363333 | 0.01060556 | 1.03 | 0.4301 |
| dose | 3 | 0.57028958 | 0.19009653 | 18.47 | <.0001 |
| type * dose | 9 | 0.16303542 | 0.01811505 | 1.76 | 0.1295 |

Tests of Hypotheses Using the Anova MS for type * block as an Error Term

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|--------|----|------------|-------------|---------|--------|
| type | 3 | 0.11087292 | 0.03695764 | 3.48 | 0.0903 |
| block | 2 | 0.07605000 | 0.03802500 | 3.59 | 0.0945 |

这是输出结果的第一部分, 给了两个方差分析的结果。第一个方差分析的结果中, type 和 block 的检验结果是不正确的, 因为它计算 F 值时采用的误差项是子区的误差, 不是全区自身的误差, type * block 的检验结果没有意义, 因为它是全区的误差。在第二个方差分析结果中, 给出了以全区误差 (type * block) 作为误差项计算而得的 type 和 block 对应的 F 值和 P 值。综合两部分方差分析结果可以看出: 因素 type、block 和交互项 type * dose 是无统计学意义的, 只有因素 dose 有统计学意义, 即不同浓度的蛇毒对小鼠抑瘤效果的影响之间存在差异。

程序 sastjfx10_9 的输出结果中还给出了两个实验因素 type 和 dose 各水平全面组合而成的 16 个实验组瘤重均数两两比较的结果。由于篇幅限制, 此处不再给出。

Fit Statistics

| | |
|--------------------------|-------|
| -2 Res Log Likelihood | -35.3 |
| AIC (smaller is better) | -29.3 |
| AICC (smaller is better) | -28.4 |
| BIC (smaller is better) | -32.0 |

Type 3 Tests of Fixed Effects

| Effect | Num DF | Den DF | F Value | Pr > F |
|-------------|--------|--------|---------|--------|
| type | 3 | 6 | 3.48 | 0.0903 |
| dose | 3 | 24 | 18.47 | <.0001 |
| type * dose | 9 | 24 | 1.76 | 0.1295 |

这是输出结果的第三部分, 是采用混合模型进行统计分析的结果。首先给出了模型拟合的有关信息: AIC = -29.3, AICC = -28.4, BIC = -32.0 等, 这些信息在分析重复测量资料时是有用的。随后给出了方差分析的结果, 这与 glm 过程计算结果相同, 此处不再赘述。

另外, mixed 过程还给出了各种实验条件下瘤重均值两两比较的结果, 其中 type 和 dose 两列表示一种实验条件, _type 和 _dose 两列表示另一种实验条件。因结果占用篇幅较大, 此处不再列出。

10.10 正交设计一元定量资料方差分析

10.10.1 问题与数据

【例 10-10】 某研究者欲确定氧化葡萄糖的最优制备条件，采用正交实验考察 pH 值、反应温度、搅拌速度 3 个实验因素的影响，因素水平见表 10-10，以生成物的醛基含量和反应时间为参考指标进行综合评分，评分越高说明氧化葡萄糖制备效率越高。实验设计及实验结果见表 10-11，请进行适当的统计分析。

表 10-10 因素水平

| 因素水平 | 反应液 pH 值 (A) | 搅拌速度 (r · min ⁻¹) (B) | 反应温度 (°C) (C) |
|------|--------------|-----------------------------------|---------------|
| 1 | 3.6 | 1000 | 42 |
| 2 | 4.5 | 1200 | 38 |
| 3 | 6.1 | 1500 | 25 |

表 10-11 正交实验设计及实验结果

| 实验号 | 1 (A) | 2 (B) | 3 | 4 (C) | 醛基含量 (mmol/g) | 反应时间 (h) | 评分 |
|-----|-------|-------|-----|-------|---------------|----------|------|
| 1 | 1 | 1 | 1 | 1 | 8.9 | 0.9 | 16.9 |
| 2 | 1 | 2 | 2 | 2 | 7.5 | 1.3 | 13.7 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 9 | 3 | 3 | 2 | 1 | 6.2 | 7.4 | 5.0 |

注：评分 = 醛基含量 × 2 - 反应时间。

【例 10-11】 某研究者欲研究采用乙醇回流法从黄芪、女贞子中提取黄芪甲甙的制备工艺，以黄芪甲甙含量为指标，采用薄层扫描法进行含量测定，研究者采用正交设计考察乙醇浓度、加乙醇量、回流时间、回流次数 4 个 3 水平的实验因素对黄芪甲甙含量的影响，在 L₉(3⁴) 正交表确定的 9 个实验点上各进行 4 次独立重复实验，因素水平见表 10-12。实验设计及实验结果见表 10-13。黄芪甲甙含量越高，效果越好。请进行适当的统计分析。

表 10-12 因素水平

| 因素水平 | 乙醇浓度 (%) (A) | 加乙醇量 (倍) (B) | 回流时间 (h) (C) | 回流次数 (次) (D) |
|------|--------------|--------------|--------------|--------------|
| 1 | 60 | 8 | 2 | 3 |
| 2 | 70 | 7 | 1.5 | 2 |
| 3 | 80 | 6 | 1 | 1 |

表 10-13 正交实验设计及实验结果

| 实验号 | 1 (A) | 2 (B) | 3 (C) | 4 (D) | 黄芪甲甙含量 (mg/mL) | | | |
|-----|-------|-------|-------|-------|----------------|------|------|------|
| | | | | | 1 | 2 | 3 | 4 |
| 1 | 1 | 1 | 1 | 1 | 1.60 | 1.52 | 1.46 | 1.68 |
| 2 | 1 | 2 | 2 | 2 | 1.50 | 1.58 | 1.60 | 1.66 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 9 | 3 | 3 | 2 | 1 | 1.53 | 1.51 | 1.43 | 1.58 |

10.10.2 对数据结构的分析

对于例 10-10 资料，研究者采用正交设计来确定氧化葡萄糖的最优制备条件，选用的是

$L_9(3^4)$ 正交表, 表头设计时在第 1、2、4 列分别安排 A、B、C 三个 3 水平的因素, 以第 3 列为空白列来估计误差, 观测指标为“评分”。

对于例 10-11 资料, 研究者采用正交设计来确定乙醇回流法提取黄芪甲甙的最佳制备工艺, 选用的是 $L_9(3^4)$ 正交表, 表头设计时在第 1、2、3、4 列分别安排 A、B、C、D 四个 3 水平的因素, 每个实验点做 4 次独立重复实验, 以估计误差, 观察指标为“黄芪甲甙含量”。

10.10.3 分析目的与统计分析方法的选择

对这两个资料, 均应采用正交设计一元定量资料方差分析处理。

10.10.4 SAS 程序

分析例 10-10 资料, SAS 程序名为 sastjfx10_10.sas。

```
data sastjfx10_10;          /* 1 */
  input a 3 b 4 c 5 y;
  cards;
  111 16.9
  122 13.7
  . . . .
  331 5.0
  ;
run;

ods html;
proc anova;                 /* 2 */
  class a b c;
  model y = a b c;
  means a b c;
run;
ods html close;
```

程序第 1 步建立数据集, a、b、c 分别代表“反应液 pH 值”、“搅拌速度”和“反应温度”, y 代表观测指标“评分”。需要注意的是, a、b、c 三个因素数据的录入格式, 在 input 语句中分别指定 a、b、c 占据第 3、4、5 列(录入数据时缩进了两列, 若不缩进则 a、b、c 分别占据第 1、2、3 列), 故在 cards 语句后录入数据时, 三个变量的取值是连着写的。当然, 由于指定了变量数据的位置, 也就无法在 input 语句后加上“@@”来实现数据的连续录入了。第 2 步调用 anova 过程对资料进行正交设计定量资料方差分析。

分析例 10-11 资料, SAS 程序名为 sastjfx10_11.sas。

```
data sastjfx10_11;          /* 1 */
  input a b c d @@ ;
  doi = 1 to 4;
  input y @@ ;
  output; end;
  cards;
  1 1 1 1 1.60 1.52 1.46 1.68
  1 2 2 2 1.50 1.58 1.60 1.66
  . . . . .
  3 3 2 1 1.53 1.51 1.43 1.58
  ;
run;

ods html;
proc anova;                 /* 2 */
  class a b c d;
  model y = a b c d;
  means a b c d;
run;
ods html close;
```

程序第 1 步建立数据集, a、b、c、d 分别代表“乙醇浓度”、“加乙醇量”、“回流时间”和“回流次数”, y 代表观测指标“黄芪甲甙含量”。do 循环语句用来实现对每个实验条件下获得的多次独立重复实验数据的录入。与程序 sastjfx10_10 不同的是, 此程序没有指定 a、b、c、d 四个因素的录入位置, 故可在 input 语句后加上“@@”来实现数据的连续录入了。第 2 步调用 anova 过程对资料进行正交设计定量资料方差分析。

10.10.5 主要分析结果及解释

以下是 SAS 软件分析例 10-10 资料的结果,即程序 sastjfx10_10.sas 输出的结果。

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|--------|----|-------------|-------------|---------|--------|
| a | 2 | 185.1622222 | 92.5811111 | 91.87 | 0.0108 |
| b | 2 | 1.6955556 | 0.8477778 | 0.84 | 0.5431 |
| c | 2 | 44.1155556 | 22.0577778 | 21.89 | 0.0437 |

这是输出结果的第一部分。可以看出,因素 a、c 各水平对观测指标的影响之间的差异有统计学意义,而因素 b 各水平对观测指标的影响之间的差异则无统计学意义。

| Level of | N | y |
|----------|---|-----------------------|
| a | | Mean Std Dev |
| 1 | 3 | 13.4000000 3.65923489 |
| 2 | 3 | 8.2666667 2.08166600 |
| 3 | 3 | 2.3000000 2.48797106 |

| Level of | N | y |
|----------|---|-----------------------|
| a | | Mean Std Dev |
| 1 | 3 | 8.43333333 7.71513664 |
| 2 | 3 | 8.13333333 7.12764571 |
| 3 | 3 | 7.40000000 2.30651252 |

| Level of | N | y |
|----------|---|------------------------|
| a | | Mean Std Dev |
| 1 | 3 | 10.83333333 5.95343038 |
| 2 | 3 | 7.70000000 5.95063022 |
| 3 | 3 | 5.43333333 4.85626743 |

这是输出结果的第二部分,给出了 a、b、c 三个因素各水平观测指标的均值和标准差。

由于因素 b 各水平对观测指标的影响之间的差异无统计学意义,可在 SAS 程序的 class、model 语句中将因素 b 删除,重新运行上述程序,得结果如下:

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|--------|----|-------------|-------------|---------|--------|
| a | 2 | 185.1622222 | 92.5811111 | 99.79 | 0.0004 |
| c | 2 | 44.1155556 | 22.0577778 | 23.77 | 0.0060 |

可以看出,因素 a、c 所对应的 P 值变得更小了。所以,在此项研究中,应重点考察因素 a 和 c,适当照顾因素 b 即可。参考输出结果的第二部分,比较因素 a、b、c 各水平的均值,由于“评分”越高越好,故选择三个因素各自均值最大的水平,分别为 a_1 、 b_1 、 c_1 ,由于研究者没有考察因素间的交互作用或专业上认为各因素间的交互作用可以忽略不计,就目前的结果而言, $a_1b_1c_1$ 的组合是氧化葡萄糖的最优制备条件,即反应液 pH 值为 3.6,搅拌速度为 $1000\text{r} \cdot \text{min}^{-1}$,反应温度为 42°C 时,氧化葡萄糖的制备效率最高。

以下是 SAS 软件分析例 10-11 资料的结果,即程序 sastjfx10_11.sas 输出的结果。

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|--------|----|----------|-------------|---------|--------|
|--------|----|----------|-------------|---------|--------|

| | | | | | |
|---|---|------------|------------|--------|--------|
| a | 2 | 0.06237222 | 0.03118611 | 6.72 | 0.0043 |
| b | 2 | 0.02302222 | 0.01151111 | 2.48 | 0.1026 |
| c | 2 | 0.08175556 | 0.04087778 | 8.81 | 0.0011 |
| d | 2 | 2.63775556 | 1.31887778 | 284.14 | <.0001 |

这是输出结果的第一部分。可以看出，因素 a、c、d 各水平对观测指标的影响之间的差异有统计学意义，而因素 b 各水平对观测指标的影响之间的差异则无统计学意义。

| Level of | N | y | |
|----------|----|------------|------------|
| a | | Mean | Std Dev |
| 1 | 12 | 1.40166667 | 0.26706599 |
| 2 | 12 | 1.30000000 | 0.26454077 |
| 3 | 12 | 1.34416667 | 0.34555511 |

| Level of | N | y | |
|----------|----|------------|------------|
| b | | Mean | Std Dev |
| 1 | 12 | 1.38416667 | 0.32444241 |
| 2 | 12 | 1.33416667 | 0.33568000 |
| 3 | 12 | 1.32750000 | 0.21528522 |

| Level of | N | y | |
|----------|----|------------|------------|
| c | | Mean | Std Dev |
| 1 | 12 | 1.28916667 | 0.31140323 |
| 2 | 12 | 1.35083333 | 0.29916424 |
| 3 | 12 | 1.40583333 | 0.26922307 |

| Level of | N | y | |
|----------|----|------------|------------|
| d | | Mean | Std Dev |
| 1 | 12 | 1.53583333 | 0.06973434 |
| 2 | 12 | 1.54416667 | 0.11131106 |
| 3 | 12 | 0.96583333 | 0.09662094 |

这是输出结果的第二部分，给出了 a、b、c、d 四个因素各水平观测指标的均值和标准差。

由于因素 b 各水平对观测指标的影响之间的差异无统计学意义，可在 SAS 程序的 class、model 语句中将因素 b 删除，重新运行上述程序，得结果如下：

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|--------|----|------------|-------------|---------|--------|
| a | 2 | 0.06237222 | 0.03118611 | 6.10 | 0.0062 |
| c | 2 | 0.08175556 | 0.04087778 | 7.99 | 0.0017 |
| d | 2 | 2.63775556 | 1.31887778 | 257.82 | <.0001 |

可以看出，因素 a、c、d 所对应的 P 值依然小于 0.05。所以，在此项研究中，应重点考察因素 a、c、d，适当照顾因素 b 即可。参考输出结果的第二部分，比较因素 a、b、c、d 各水平的均值，由于黄芪甲甙含量越高越好，故选择四个因素各自均值最大的水平，分别为 a_1 、 b_1 、 c_3 、 d_2 ，由于研究者没有考察因素间的交互作用或专业上认为各因素间的交互作用可以忽略不计，就目前的结果而言， $a_1b_1c_3d_2$ 的组合是乙醇回流法提取黄芪甲甙的最优制备条件，即加 8 倍量浓度为 60% 的乙醇回流提取 2 次，每次回流 1h，黄芪甲甙的制备效果最好。

10.11 重复测量设计一元定量资料方差分析

10.11.1 问题与数据

【例 10-12】 某课题组研究益髓生血颗粒治疗^{-SEA}/ $\alpha\alpha^{CS}$ 基因型患者血红蛋白 H(hemoglobin H, HbH)病的临床疗效,选取 13 名^{-SEA}/ $\alpha\alpha^{CS}$ 基因型 HbH 病患者,给以益髓生血颗粒治疗,疗程为 3 个月。分别记录患者治疗前、服药 1 个月、服药 2 个月及服药 3 个月时血红蛋白的含量,结果见表 10-14。请进行合适的统计分析。

表 10-14 益髓生血颗粒治疗前后^{-SEA}/ $\alpha\alpha^{CS}$ 基因型 HbH 患者血红蛋白含量

| 患者编号 | 血红蛋白含量(g/L) | | | | |
|------|-------------|-------|---------|---------|---------|
| | 观测时间: | 治疗前 | 服药 1 个月 | 服药 2 个月 | 服药 3 个月 |
| 1 | | 94.98 | 102.33 | 98.86 | 102.31 |
| 2 | | 91.41 | 98.79 | 95.46 | 98.66 |
| ... | | ... | ... | ... | ... |
| 13 | | 79.16 | 86.63 | 83.77 | 86.12 |

【例 10-13】 随机选取 6 只狗进行试验,公、母各 3 只,服用某种中药复方制剂后在 3 个不同时间点测定血药浓度,数据见表 10-15。请进行合适的统计分析。

表 10-15 每只狗服药后 3 个不同时间点上血药浓度的测定结果

| 性别 | 狗号 | 血药浓度($\mu\text{g/mL}$) | | | |
|----|----|--------------------------|-------|-------|------|
| | | 时间(h): | 6 | 8 | 24 |
| 公 | 1 | | 169.8 | 137.2 | 31.9 |
| | 2 | | 158.4 | 133.2 | 18.7 |
| | 3 | | 142.8 | 126.6 | 18.1 |
| 母 | 4 | | 131.8 | 130.3 | 17.5 |
| | 5 | | 118.0 | 124.5 | 18.7 |
| | 6 | | 120.8 | 123.5 | 24.8 |

【例 10-14】 某研究者欲观察隔日禁食结合耐饥食饵对小鼠生理指标的影响,将 20 只成年 ICR 小鼠随机等分为对照组(自由摄食标准饲料)、食饵组(自由摄食耐饥食饵)、隔日禁食组(隔日禁食标准饲料)、辟谷食饵组(隔日禁食耐饥食饵),实验 8 周,自由饮水,观察小鼠体重的变化,见表 10-16。请进行适当的统计分析。

表 10-16 4 组小鼠体重变化

| 摄食方式 | 饲料种类 | 编号 | 体重(g) | | | | |
|------|------|-----|-------|-------|-------|-------|-------|
| | | | 时间: | 1 周 | 2 周 | 4 周 | 8 周 |
| 自由摄食 | 标准饲料 | 1 | | 30.82 | 41.18 | 41.83 | 40.46 |
| | | ... | | ... | ... | ... | ... |
| | | 5 | | 31.09 | 41.14 | 44.33 | 41.12 |
| | 耐饥食饵 | 1 | | 32.38 | 40.32 | 41.45 | 39.61 |
| | | ... | | ... | ... | ... | ... |
| 隔日禁食 | 标准饲料 | 5 | | 32.04 | 40.89 | 43.52 | 40.04 |
| | | 1 | | 29.52 | 34.15 | 34.69 | 37.43 |
| | | ... | | ... | ... | ... | ... |
| | 耐饥食饵 | 5 | | 32.89 | 35.96 | 36.65 | 41.08 |
| | | 1 | | 33.55 | 35.04 | 35.89 | 38.88 |
| | | ... | | ... | ... | ... | ... |
| | | 5 | | 31.59 | 32.21 | 33.40 | 36.70 |

【例 10-15】 某研究者用贲门癌患者的标本制成液体，在三种不同处理条件下观察鸡胚背根神经节与鸡胚交感神经节中长出突起的神经节比例。现有贲门癌患者 10 例，将每人的标本均分成三份，分别给予三种不同的处理(因素 A)，即 A_1 (加入 100ng/mL 神经生长因子)、 A_2 (加入 200ng/mL 神经生长因子)和 A_3 (单用贲门癌培养液)；并对每种处理后的标本中的两种类型的神经节(因素 B)，即 B_1 (背根神经节)与 B_2 (交感神经节)，观测长出突起的神经节的比例(Y)。实验结果见表 10-17。请进行适当的统计分析。

表 10-17 贲门癌患者的标本经三种处理后两种神经节中长出突起的神经节的比例

| 病 例 号 | Y(长出突起的神经节的比例) | | | | | |
|-------|-----------------|-------|------------|-------|------------|-------|
| | $A_1(B_1)$ | B_2 | $A_2(B_1)$ | B_2 | $A_3(B_1)$ | B_2 |
| 1 | 0.50 | 0.43 | 0.50 | 0.43 | 0.80 | 0.50 |
| 2 | 0.63 | 0.38 | 0.55 | 0.50 | 0.71 | 0.63 |
| ... | ... | ... | ... | ... | ... | ... |
| 10 | 0.44 | 0.43 | 0.45 | 0.40 | 0.60 | 0.75 |

10.11.2 对数据结构的分析

例 10-12 中，对每一个患者来说，在 4 个时间点上分别被测量血红蛋白含量，说明“时间”因素是一个重复测量因素，且所有患者均接受同一种治疗方法——服用益髓生血颗粒，因而这是具有一个重复测量因素的单因素设计一元定量资料。

例 10-13 中，对每只狗来说，在 3 个时间点上分别被测量血药浓度的数值，说明“时间”因素是一个重复测量因素。此外，还有一个实验因素“性别”，它把全部狗分成公与母两组，故“性别”也可被称为“实验分组”因素，因而这是具有一个重复测量因素的两因素设计一元定量资料。

例 10-14 中，对每一只小鼠来说，在 4 个时间点上分别被测量其体重，说明“时间”因素是一个重复测量的因素。此外，还有两个实验因素“摄食方式”(自由摄食或隔日禁食)和“饲料种类”(标准饲料或耐饥食饵)，因而这是具有一个重复测量因素的三因素设计一元定量资料。

例 10-15 中，涉及两个实验因素，即“处理”和“神经节类型”，前者有 3 个水平，后者有 2 个水平。由于是对同一个患者的标本采用不同的处理，同时在不同类型神经节中重复获得指标“长出突起的神经节的比例”的 6 个观测值，因而两个实验因素都是重复测量因素。所以该资料所对应的实验设计类型为具有两个重复测量因素的两因素设计。

10.11.3 分析目的与统计分析方法的选择

分析目的都是比较不同实验条件下定量观测指标的平均值之间的差别是否具有统计学意义。

例 10-12，应选用具有一个重复测量因素的单因素设计一元定量资料方差分析；例 10-13，应选用具有一个重复测量因素的两因素设计一元定量资料方差分析；例 10-14，应选用具有一个重复测量因素的三因素设计一元定量资料方差分析；例 10-15，应选用具有两个重复测量因素的两因素设计一元定量资料方差分析。

10.11.4 SAS 程序

对例 10-12 进行具有一个重复测量的单因素设计一元定量资料方差分析，SAS 程序名为 sastjfx10_12. sas。

```

data sastjfx10_12;          /* 1 */
  do person = 1 to 13;
    do month = 0 to 3;
      input y @@ ;
      output;
    end; end;
  cards;
  94.98  102.33  98.86  102.31
  91.41  98.79  95.46  98.66
  . . . . .
  79.16  86.63  83.77  86.12
  ;
run;
ods html;
proc mixed;                 /* 2 */
  class person month;
  model y = month;
  repeated/type = VC sub = person;
run;
proc mixed;                 /* 3 */
  class person month;
  model y = month;
  repeated/type = CS sub = person;
run;

proc mixed;                 /* 4 */
  class person month;
  model y = month;
  repeated/type = UN sub = person;
run;
proc mixed;                 /* 5 */
  class person month;
  model y = month;
  repeated/type = AR(1) sub = person;
run;
proc mixed;                 /* 6 */
  class person month;
  model y = month;
  repeated/type = SP (POW) (month)
  sub = person;
run;
ods html close;

```

程序第1步建立数据集, person 代表“患者编号”, month 代表观察时间, y 代表观测指标“血红蛋白含量”。第2、3、4、5、6步分别调用 mixed 过程, 采用 VC、CS、UN、AR(1)、SP(POW)五种协方差结构模型对资料进行方差分析。其中, SP(POW)模型在使用时要求将“repeated/type = sp(pow) (c-list)”语句中“c-list”替换成重复测量因素(可以是多个)的名称, 但这个或这些重复测量因素应为定量变量且其水平数不能太少。本例中, 重复测量因素为观测时间, 其包含4个水平(0、1、2、3), 可近似视为定量变量, 故可选用 SP(POW)模型。若某重复测量因素为定性变量(如部位等)或虽可视为定量变量但水平数较少(如小于等于3)时, 是不适宜选用此模型的。此外, 定量的重复测量因素在赋值时需保证赋值数值的间隔与实际观测时间间隔成比例, 较好且最简单的方法是以其真实水平代入。如本例, 可以0、1、2、3代表治疗前、服药1个月、服药2个月、服药3个月, 而不宜随意赋值, 如赋为1、3、5、9则不妥。

对例10-13进行具有一个重复测量因素的两因素设计一元定量资料方差分析, SAS 程序名为 sastjfx10_13.sas。

```

data sastjfx10_13;          /* 1 */
  do sex = 1 to 2;
    do dog = 1 to 3;
      do time = 1 to 3;
        input y @@ ; output;
      end; end; end;

proc mixed data = sastjfx10_13; /* 5 */
  clas ssex dog time;
  model y = sex | time;
  repeated/type = AR(1)
  sub = dog (sex);

```



```

end; end;end;
cards;
169.8 137.2 31.9 158.4 133.2 18.7
142.8 126.6 18.1 131.8 130.3 17.5
118.0 124.5 18.7 120.8 123.5 24.8
;
run;
proc mixed data=sastjfx10_13; /* 2 */
  classsex dog time;
  model y=sex|time;
  repeated/type=VC sub=dog (sex);
  ods output fitstatistics=a;
  ods output dimensions=a1;
run;
proc mixed data=sastjfx10_13; /* 3 */
  classsex dog time;
  model y=sex|time;
  repeated/type=CS sub=dog (sex);
  ods output fitstatistics=b;
  ods output dimensions=b1;
run;
proc mixed data=sastjfx10_13; /* 4 */
  classsex dog time;
  model y=sex|time;
  repeated/type=UN sub=dog (sex);
  ods output fitstatistics=c;
  ods output dimensions=c1;
run;

ods output fitstatistics=d;
ods output dimensions=d1;
run;
%MACRO SHUJU(dataset,y); /* 6 */
data &dataset;
  set &dataset;
  &y=value;
  drop value;
run;
%MEND SHUJU;
%SHUJU(a,VC) %SHUJU(b,CS)
%SHUJU(c,UN) %SHUJU(d,AR1)
%SHUJU(a1,VC) %SHUJU(b1,CS)
%SHUJU(c1,UN) %SHUJU(d1,AR1)
data e; /* 7 */
  merge a b c d;
run;
data e1; /* 8 */
  merge a1 b1 c1 d1;
run;
ods html;
proc print data=e; /* 9 */
  format_numeric_5.1;
run;
proc print data=e1; /* 10 */
run;
ods html close;

```

【程序说明】 程序第 1 步建立数据集，sex 代表“性别”，time 代表“测定时间”，dog 代表“受试动物个体”，y 代表观测指标“血药浓度”。第 2、3、4、5 步分别调用 MIXED 过程，采用 VC、CS、UN、AR(1) 四种协方差结构模型对资料进行方差分析。第 6 步建立宏 SHUJU，以实现对数据集中已有变量 value 的更名。第 7、8 步均用来实现对不同数据集的横向合并。第 9、10 步均用来将数据集中的内容输出到 output 窗口中去。第 2、3、4、5 步所用语句基本相同，仅在“type=”后的选项不同，4 个过程分别指定了 4 种协方差结构模型。MIXED 过程中 repeated 语句用来规定个体的重复测量的协方差结构，“/”后的 sub(也可写为 subject)用来指定数据集中的个体，若不含分组因素，直接在“sub=”后面给出受试对象个体变量名称即可；若含有分组因素，则在“sub=”后面给出受试对象个体变量名称的同时，还需在后面加注“()”，括号内填入分组变量名称。在调用 mixed 过程进行方差分析时，使用了两个 ods(output delivery system) 语句，分别用来将模型拟合的有关信息(fitstatistics)和模型维度有关参数(dimensions) 输出。

对例 10-14 进行具有一个重复测量的三因素设计一元定量资料方差分析，SAS 程序名为 sastjfx10_14.sas。

```

data sastjfx10_14; /* 1 */
  do style=1 to 2;
  do forage=1 to 2;
  do subject=1 to 5;
  do time=1,2,4,8;
  input y @@ ;
  output;
  end; end; end; end;
  cards;
  30.82 41.18 41.83 40.46 29.03
  35.11 38.86 38.84 31.22 39.40
  . . . . .
  39.62 31.59 32.21 33.40 36.70
;
run;
proc mixed data=sastjfx10_14; /* 2 */
  class style forage subject time;
  model y=style|forage|time;
  repeated/type=VC sub=subject(style
forage);
  ods output fitstatistics=a;
  ods output dimensions=a1;
run;
proc mixed data=sastjfx10_14; /* 3 */
  class style forage subject time;
  model y=style|forage|time;
  repeated/type=CS sub=subject(style
forage);
  ods output fitstatistics=b;
  ods output dimensions=b1;
run;
proc mixed data=sastjfx10_14; /* 4 */
  class style forage subject time;
  model y=style|forage|time;
  repeated/type=UN sub=subject(style
forage);
  ods output fitstatistics=c;
  ods output dimensions=c1;
run;
proc mixed data=sastjfx10_14; /* 5 */
  class style forage subject time;
  model y=style|forage|time;
  repeated/type=AR(1)
  sub=subject(style forage);
  ods output fitstatistics=d;
  ods output dimensions=d1;
run;
proc mixed data=sastjfx10_14; /* 6 */
  class style forage subject time;
  model y=style|forage|time;
  repeated/type=SP(POW)(time)
  sub=subject(style forage);
  ods output fitstatistics=e;
  ods output dimensions=e1;
run;
%macro shuju(dataset,y); /* 7 */
  data &dataset;
  set &dataset;
  rename value=&y; run;
%mend shuju;
%shuju(a,VC) %shuju(b,CS)
%shuju(c,UN) %shuju(d,AR1)
%shuju(e,SP) %shuju(a1,VC)
%shuju(b1,CS) %shuju(c1,UN)
%shuju(d1,AR1) %shuju(e1,SP)
data f; /* 8 */
  merge a b c d e;
run;
data f1; /* 9 */
  merge a1 b1 c1 d1 e1;
run;
ods html;
proc print data=f; /* 10 */
  format_numeric_5.1;
run;
proc print data=f1; /* 11 */
run;
ods html close;

```

程序第1步建立数据集，style代表“摄食方式”，forage代表“饲料种类”，time代表“时间”，subject代表“小鼠个体”，y代表观测指标“体重”。第2、3、4、5、6步分别调用mixed过程，采用VC、CS、UN、AR(1)、SP(POW)五种协方差结构模型对资料进行方差分析，其中model语句后“style|forage|time”相当于“style forage time style * forage style * time forage * time style * forage * time”，即这三个因素的主效应及各级交互作用的效应。第7步建立宏shuju，以实现对数据集中已有变量value的更名。第8、9步均用来实现对不同数据集的横向合并。第10、11步均用来将数据集中的内容输出到output窗口中去。第2、3、4、5、6步分别调用mixed

过程,几个过程所用语句基本相同,仅在“type =”后的选项不同,5 个过程分别指定了 5 种协方差结构模型。mixed 过程中 repeated 语句用来规定个体的重复测量的协方差结构,“/”后的 sub(也可写为 subject)用来指定数据集中的个体,若不含分组因素,直接在“sub =”后面给出受试对象个体变量名称即可,如程序 sastjfx10_12;若含有分组因素,则在“sub =”后面给出受试对象个体变量名称的同时,还需在后面加注“()”,括号内填入分组变量名称,如程序 sastjfx10_14。在调用 mixed 过程进行方差分析时,使用了两个 ods 语句,分别用来将模型拟合的有关信息(fitstatistics)和模型维度有关参数(dimensions)输出。

对例 10-15 进行具有两个重复测量因素的两因素设计一元定量资料方差分析,SAS 程序名为 sastjfx10_15.sas。

```

data sastjfx10_15;          /* 1 */
  do patient = 1 to 10;
  do treat = 1 to 3;
  do type = 1 to 2;
  input y @@ ;
  output;
  end; end; end;
  cards;
  0.50 0.43 0.50 0.43 0.80 0.50
  0.63 0.38 0.55 0.50 0.71 0.63
  . . . . .
  0.44 0.43 0.45 0.40 0.60 0.75
  ;
run;
proc mixed data=sastjfx10_15; /* 2 */
  class patient treat type;
  model y=treat|type;
  repeated/type=VC sub=patient;
  ods output fitstatistics=a;
  ods output dimensions=a1;
run;
proc mixed data=sastjfx10_15; /* 3 */
  class patient treat type;
  model y=treat|type;
  repeated/type=CS sub=patient;
  ods output fitstatistics=b;
  ods output dimensions=b1;
run;
proc mixed data=sastjfx10_15; /* 4 */
  class patient treat type;
  model y=treat|type;
  repeated/type=UN sub=patient;
  ods output fitstatistics=c;
  ods output dimensions=c1;
run;
proc mixed data=sastjfx10_15; /* 5 */
  class patient treat type;
  model y=treat|type;
  repeated/type=AR(1) sub=patient;
  ods output fitstatistics=d;
  ods output dimensions=d1;
run;
ods html;
proc sql;                  /* 6 */
  select a.Descr,a.VALUE as
  VC,b.VALUE as CS,c.VALUE as
  UN,d.VALUE as AR1
  from a,b,c,d
  where
  a.Descr=b.Descr=
  c.Descr=d.Descr;
quit;
proc sql;                  /* 7 */
  select a1.Descr,a1.VALUE as
  VC,b1.VALUE as CS,c1.VALUE as
  UN,d1.VALUE as AR1
  from a1,b1,c1,d1
  where a1.Descr=b1.Descr=c1.Descr=
  d1.Descr;
quit;
ods html close;

```

程序第 1 步建立数据集,patient 代表“病例号”,treat 代表“处理”,type 代表“神经节类型”,y 代表观测指标“长出突起的神经节的比例”。第 2、3、4、5 步分别调用 mixed 过程,采用 VC、CS、UN、AR(1)四种协方差结构模型对资料进行方差分析。第 6、7 步将不同数据集的内容进行合并查询,这与程序 sastjfx10_14 中的宏 shuju 功能相同。

10.11.5 主要分析结果及解释

以下是对例 10-12 的分析结果,即程序 sastjfx10_12.sas 的部分输出结果。

| Fit Statistics | | | | |
|-------------------------------|--------------------------|--------|---------|--------|
| | - 2 Res Log Likelihood | | 384.1 | |
| | AIC(smaller is better) | | 386.1 | |
| | AICC(smaller is better) | | 386.2 | |
| | BIC(smaller is better) | | 386.7 | |
| Type 3 Tests of Fixed Effects | | | | |
| Effect | Num DF | Den DF | F Value | Pr > F |
| month | 3 | 36 | 1.07 | 0.3734 |

这是采用 VC(方差分量型) 协方差结构模型进行混合效应模型分析的输出结果。

| Fit Statistics | | | | |
|-------------------------------|--------------------------|--------|---------|---------|
| | - 2 Res Log Likelihood | | 146.0 | |
| | AIC(smaller is better) | | 150.0 | |
| | AICC(smaller is better) | | 150.3 | |
| | BIC(smaller is better) | | 151.1 | |
| Type 3 Tests of Fixed Effects | | | | |
| Effect | Num DF | Den DF | F Value | Pr > F |
| month | 3 | 36 | 1267.75 | < .0001 |

这是采用 CS(复合对称型) 协方差结构模型进行混合效应模型分析的输出结果。

| Iteration History | | | |
|-------------------|-------------|------------------|------------|
| Iteration | Evaluations | - 2 Res Log Like | Criterion |
| 0 | 1 | 384.11886968 | |
| 1 | 1 | - 210.88040371 | 0.00000214 |
| ... | ... | ... | ... |
| 11 | 0 | - 210.88040371 | 0.00000214 |

WARNING: Did not converge.

这是采用 UN(无结构型) 协方差结构模型进行混合效应模型分析的输出结果。

| Fit Statistics | | | | |
|-------------------------------|--------------------------|--------|---------|---------|
| | - 2 Res Log Likelihood | | 163.6 | |
| | AIC(smaller is better) | | 167.6 | |
| | AICC(smaller is better) | | 167.9 | |
| | BIC(smaller is better) | | 168.8 | |
| Type 3 Tests of Fixed Effects | | | | |
| Effect | Num DF | Den DF | F Value | Pr > F |
| month | 3 | 36 | 990.63 | < .0001 |

这是采用 AR(1)(一阶自回归型) 协方差结构模型进行混合效应模型分析的输出结果。

| Fit Statistics | |
|------------------------|-------|
| - 2 Res Log Likelihood | 163.6 |

| | |
|--------------------------|-------|
| AIC(smaller is better) | 167.6 |
| AICC(smaller is better) | 167.9 |
| BIC(smaller is better) | 168.8 |

Type 3 Tests of Fixed Effects

| Effect | Num DF | Den DF | F Value | Pr > F |
|--------|--------|--------|---------|---------|
| month | 3 | 36 | 990.63 | < .0001 |

这是采用 SP(POW)(空间幂相关型)协方差结构模型进行混合效应模型分析的输出结果。

以上是采用 mixed 过程进行混合效应模型分析的结果。这里给出了模型的拟合信息和固定效应假设检验的结果。通常关注的是固定效应假设检验的结果。但从上面的结果中可以看出,采用不同的协方差结构模型得出的固定效应假设检验的结果是不完全相同的,其中 glm 过程计算出来的结果与 CS 协方差结构模型(采用 REML 即约束最大似然估计法)得出的结果是相同的,它是最简单的模型。那么,在采用这 5 种协方差结构模型计算所得的结果中应以哪个结果为准呢?这就是模型选择的问题了。

通常情况下,可以 Akaike 的信息准则 AIC 值或 Schwarz 的信息准则 BIC(或叫 SBC)值来选择协方差结构模型。AIC 值和 BIC 值越小,协方差结构模型拟合越好;若两个协方差结构模型拟合的 AIC 值和 BIC 值接近,还可参考 $-2\log L(-2 \text{ Res Log Likelihood})$ 的数值,小者为优。当两个协方差结构模型分别包含 $q+v$ 和 q 个参数时,可用这两个模型的 -2 倍的对数似然函数值构造出似然比统计量,采用 χ^2 检验进行推断,见式(10-1)。若似然比统计量对应的 P 值大于设定的临界值,则两个协方差结构模型对资料的拟合效果之间的差异无统计学意义,此时可选择参数个数较少的那个协方差结构模型。若似然比统计量对应的 P 值小于设定的临界值,则两个协方差结构模型对资料的拟合效果之间的差异有统计学意义,此时可选择模型中参数较多[-2 倍的对数似然函数值($-2\log L$)较小]的那个协方差结构模型。

$$\chi_v^2 = -2\log L_q - (-2\log L_{q+v}) \quad (10-1)$$

式中, χ_v^2 服从自由度为 v 的 χ^2 分布, $-2\log L_q$ 和 $-2\log L_{q+v}$ 分别为含 q 和 $q+v$ 个协方差参数模型的 -2 倍的对数似然函数值。

现将本资料输出结果的模型拟合信息部分的主要内容汇总在表 10-18 中,以利于比较。

表 10-18 用 5 种类型的协方差结构模型拟合本资料的拟合效果比较

| 判断准则 | 类型: | VC | CS | UN | AR(1) | SP(POW) |
|------------|-----|-------|-------|----|-------|---------|
| AIC | | 386.1 | 150.0 | — | 167.6 | 167.6 |
| BIC | | 386.7 | 151.1 | — | 168.8 | 168.8 |
| $-2\log L$ | | 384.1 | 146.0 | — | 163.6 | 163.6 |

注:各模型中待估计的协方差结构中参数的个数依次为:1、2、10、2、2。

由上面的结果可知:采用 UN(无结构型)协方差结构模型拟合本资料,其迭代过程并没有收敛,拟合效果不好。对比其他四种协方差结构模型的拟合情况,可以看出采用 CS 协方差结构模型拟合本资料效果最好,因为评价拟合效果的三个准则中 CS 模型对应的值均为最小且其 $-2\log L$ 值较参数个数比其少的模型 VC 的值差异也比较大。所以此时应以 CS(复合对称型)协方差结构模型进行混合模型分析的输出结果为准,其中 $F = 1267.75$, $P < 0.0001$ 。说明不同

时间点上测得的 HbH 患者血红蛋白含量均值之间的差异存在统计学意义。欲得出各时间点上患者血红蛋白含量均值之间两两比较的结果,可将原程序中所有过程步删除,在 data 步后换上以下过程步:

```
ods html;
proc mixed;
  class person month;
  model y = month / ddfm = sat;
  repeated / type = CS sub = person;
  lsmeans month / pdiff;
run;
ods html close;
```

其中,“ddfm = sat”说明采用 Satterthwaite 近似的方法来计算分母的自由度,从而得到一个更加准确的近似 F 值。但是,此法计算的结果有时会与 glm 过程计算的结果有所不同。

程序运行后,得结果如下:

| Least Squares Means | | | | | | |
|---------------------|-------|----------|----------------|----|---------|---------|
| Effect | month | Estimate | Standard Error | DF | t Value | Pr > t |
| month | 0 | 723.9385 | 3.2968 | 12 | 24.25 | <.0001 |
| month | 1 | 87.4023 | 3.2968 | 12 | 26.51 | <.0001 |
| month | 2 | 84.5138 | 3.2968 | 12 | 25.64 | <.0001 |
| month | 3 | 86.9200 | 3.2968 | 12 | 26.36 | <.0001 |

| Differences of Least Squares Means | | | | | | |
|------------------------------------|-------|--------|----------|----------------|----|---------|
| Effect | month | _month | Estimate | Standard Error | DF | Pr > t |
| month | 0 | 1 | -7.4638 | 0.1355 | 36 | <.0001 |
| month | 0 | 2 | -4.5754 | 0.1355 | 36 | <.0001 |
| month | 0 | 3 | -6.9815 | 0.1355 | 36 | <.0001 |
| month | 1 | 2 | 2.8885 | 0.1355 | 36 | <.0001 |
| month | 1 | 3 | 0.4823 | 0.1355 | 36 | 0.0011 |
| month | 2 | 3 | -2.4062 | 0.1355 | 36 | <.0001 |

这是各时间点上患者血红蛋白含量均值之间两两比较的结果,首先给出了各时间点患者血红蛋白含量均值与 0 比较的检验结果,没有太大实际意义。后面给出了两两比较的结果,month 和 _month 列中的 0、1、2、3 分别代表时间的 4 个水平,即治疗前、服药 1 个月、服药 2 个月、服药 3 个月。读者可查看最后两列给出的两两比较的结果,可发现这 4 个时间点上患者血红蛋白含量均值之间的差异均有统计学意义,说明 4 个时间点上患者血红蛋白含量各不相同,根据上面给出的均值大小可判断,各时间点患者血红蛋白含量的大小关系为:服药 1 个月 > 服药 3 个月 > 服药 2 个月 > 治疗前。

以下是对例 10-13 的分析结果,即程序 sastjfx10_13.sas 的输出结果。

| Obs | Descr | VC | CS | UN | AR1 |
|-----|--------------------------|------|------|------|------|
| 1 | -2 Res Log Likelihood | 89.6 | 86.1 | 68.0 | 85.8 |
| 2 | AIC(smaller is better) | 91.6 | 90.1 | 80.0 | 89.8 |
| 3 | AICC(smaller is better) | 92.0 | 91.5 | 96.8 | 91.1 |
| 4 | BIC(smaller is better) | 91.4 | 89.7 | 78.8 | 89.4 |

| Obs | Descr | VC | CS | UN | AR1 |
|-----|-----------------------|----|----|----|-----|
| 1 | Covariance Parameters | 1 | 2 | 6 | 2 |
| 2 | Columns in X | 12 | 12 | 12 | 12 |
| 3 | Columns in Z | 0 | 0 | 0 | 0 |
| 4 | Subjects | 6 | 6 | 6 | 6 |
| 5 | Max Obs Per Subject | 3 | 3 | 3 | 3 |

这是上述程序中 ods 输出的结果。首先给出了 4 种协方差结构模型拟合本资料的有关情况, 然后给出了协方差阵的有关信息 (Covariance Parameters 表示模型中待估计的协方差结构中参数的个数)。采用不同的协方差结构模型得出的固定效应假设检验的结果是不完全相同的, 那么, 在采用这 4 种协方差结构模型计算所得的结果中应以哪个结果为准呢? 这就是模型选择的问题了。

通常情况下, 可以 Akaike 的信息准则 AIC 值或 Schwarz 的信息准则 BIC (或叫 SBC) 值来选择协方差结构模型。AIC 值和 BIC 值越小, 协方差结构模型拟合越好; 若两个协方差结构模型拟合的 AIC 值和 BIC 值接近, 还可参考 $-2\log L(-2 \text{ Res Log Likelihood})$ 的数值, 小者为优。当两个协方差结构模型分别包含 $q+v$ 和 q 个参数时, 可用这两个 -2 倍的对数似然函数值构造出似然比统计量, 采用 χ^2 检验进行推断, 见式 (10-1)。若似然比统计量对应的 P 值大于设定的临界值, 则两个协方差结构模型对资料的拟合效果之间的差异无统计学意义, 此时可选择参数个数较少的那个协方差结构模型。若似然比统计量对应的 P 值小于设定的临界值, 则两个协方差结构模型对资料的拟合效果之间的差异有统计学意义, 此时应选择参数较多的那个协方差结构模型。

此例比较 4 种模型拟合资料情况的 AIC、BIC 数值, 可发现 UN 与 AR(1) 两种协方差结构模型拟合资料较好。现比较这两种协方差结构模型拟合资料的效果之间的差异是否有统计学意义。

$$\chi^2_v = -2\log L_q - (-2\log L_{q+v}) = 85.8 - 68.0 = 17.8$$

由 ods 输出结果的第二部分可知, $q=2$, $q+v=6$, 所以 $v=4$ 。而 $\chi^2_{0.05(4)} = 9.49 < 17.8$, 故 $P < 0.05$ 。可认为不适合用 AR(1) 模型取代 UN 模型, 所以最后的结论应按 UN 协方差结构模型计算出来的结果来下。其假设检验结果为:

| Type 3 Tests of Fixed Effects | | | | |
|-------------------------------|--------|--------|---------|--------|
| Effect | Num DF | Den DF | F Value | Pr > F |
| sex | 1 | 4 | 7.16 | 0.0555 |
| time | 2 | 4 | 973.60 | <.0001 |
| sex * time | 2 | 4 | 13.46 | 0.0167 |

由上述结果可知: 测定时间 (time)、性别与测定时间的交互作用 (sex * time) 均有统计学意义。

欲得出各均数之间两两比较的结果, 可将原程序中所有过程步删除, 换上以下过程步:

```
ods html;
proc mixed data = sastjfx10_13.sas;
  class sex dog time;
  model y = sex | time / ddfm = sat;
  repeated / type = UN sub = dog (sex);
  lsmeans sex * time / diff;
run;
ods html close;
```

程序运行后,得如下结果:

| Least Squares Means | | | | | | | | | |
|---------------------|-----|------|----------|----------------|----|--------|---------|--|--|
| Effect | sex | time | Estimate | Standard Error | DF | tValue | Pr > t | | |
| sex * time | 1 | 1 | 157.00 | 6.2840 | 4 | 24.98 | <.0001 | | |
| sex * time | 1 | 2 | 132.33 | 2.6499 | 4 | 49.94 | <.0001 | | |
| sex * time | 1 | 3 | 22.9000 | 3.5628 | 4 | 6.43 | 0.0030 | | |
| sex * time | 2 | 1 | 123.53 | 6.2840 | 4 | 19.66 | <.0001 | | |
| sex * time | 2 | 2 | 126.10 | 2.6499 | 4 | 47.59 | <.0001 | | |
| sex * time | 2 | 3 | 20.3333 | 3.5628 | 4 | 5.71 | 0.0047 | | |

| Differences of Least Squares Means | | | | | | | | | |
|------------------------------------|-----|------|------|-------|----------|----------------|------|--------|---------|
| Effect | sex | time | _sex | _time | Estimate | Standard Error | DF | tValue | Pr > t |
| sex * time | 1 | 1 | 1 | 2 | 24.6667 | 3.7298 | 4 | 6.61 | 0.0027 |
| sex * time | 1 | 1 | 1 | 3 | 134.10 | 5.1995 | 4 | 25.79 | <.0001 |
| sex * time | 1 | 1 | 2 | 1 | 33.4667 | 8.8869 | 4 | 3.77 | 0.0197 |
| sex * time | 1 | 1 | 2 | 2 | 30.9000 | 6.8199 | 4.04 | 4.53 | 0.0103 |
| sex * time | 1 | 1 | 2 | 3 | 136.67 | 7.2238 | 5.35 | 18.92 | <.0001 |
| sex * time | 1 | 2 | 1 | 3 | 109.43 | 3.4525 | 4 | 31.70 | <.0001 |
| sex * time | 1 | 2 | 2 | 1 | 8.8000 | 6.8199 | 4.04 | 1.29 | 0.2658 |
| sex * time | 1 | 2 | 2 | 2 | 6.2333 | 3.7476 | 4 | 1.66 | 0.1716 |
| sex * time | 1 | 2 | 2 | 3 | 112.00 | 4.4403 | 6.46 | 25.22 | <.0001 |
| sex * time | 1 | 3 | 2 | 1 | -100.63 | 7.2238 | 5.35 | -13.93 | <.0001 |
| sex * time | 1 | 3 | 2 | 2 | -103.20 | 4.4403 | 6.46 | -23.24 | <.0001 |
| sex * time | 1 | 3 | 2 | 3 | 2.5667 | 5.0386 | 4 | 0.51 | 0.6373 |
| sex * time | 2 | 1 | 2 | 2 | -2.5667 | 3.7298 | 4 | -0.69 | 0.5292 |
| sex * time | 2 | 1 | 2 | 3 | 103.20 | 5.1995 | 4 | 19.85 | <.0001 |
| sex * time | 2 | 2 | 2 | 3 | 105.77 | 3.4525 | 4 | 30.64 | <.0001 |

这是两种性别的受试对象各时间点上血药浓度均值之间两两比较的结果,首先给出了每种性别受试对象各时间点上血药浓度均值与0比较的检验结果,没有任何实际意义。后面给出了两两比较的结果,sex和time列指定其中的某性别受试对象某个时间点上的数据,_sex和_time列也指定其中的某性别受试对象某个时间点上的数据。公狗服药后8h血药浓度与母狗服药后6h、8h血药浓度均值之间的差异无统计学意义($P=0.2658$ 、 $P=0.1716$);公狗服药后24h血药浓度与母狗服药后24h血药浓度均值之间的差异无统计学意义($P=0.6373$);母狗服药后6h血药浓度与服药后8h血药浓度均值之间的差异无统计学意义($P=0.5292$);其余服药后血药浓度均值之间的比较均有统计学意义。

以下是对例10-14的分析结果,即程序sastjfx10_14.sas的输出结果。

| Obs | Descr | VC | CS | UN | AR1 | SP |
|-----|--------------------------|-------|-------|-------|-------|-------|
| 1 | -2 Res Log Likelihood | 301.6 | 237.2 | 219.1 | 235.5 | 241.7 |
| 2 | AIC(smaller is better) | 303.6 | 241.2 | 239.1 | 239.5 | 245.7 |
| 3 | AICC(smaller is better) | 303.7 | 241.3 | 243.3 | 239.7 | 245.9 |
| 4 | BIC(smaller is better) | 304.6 | 243.1 | 249.1 | 241.5 | 247.7 |
| Obs | Descr | VC | CS | UN | AR1 | SP |

| | | | | | | |
|---|-----------------------|----|----|----|----|----|
| 1 | Covariance Parameters | 1 | 2 | 10 | 2 | 2 |
| 2 | Columns in X | 45 | 45 | 45 | 45 | 45 |
| 3 | Columns in Z | 0 | 0 | 0 | 0 | 0 |
| 4 | Subjects | 20 | 20 | 20 | 20 | 20 |
| 5 | Max Obs Per Subject | 4 | 4 | 4 | 4 | 4 |

这是上述程序中 ods(output delivery system)输出的结果。首先给出了 5 种协方差结构模型拟合本资料的有关情况, 然后给出了协方差阵的有关信息(Covariance Parameters 表示模型中待估计的协方差结构中参数的个数)。比较 5 种模型拟合资料情况的 AIC、BIC 数值, 可发现 UN 与 AR(1) 两种协方差阵模型拟合资料较好。现比较这两种协方差阵模型拟合资料的效果之间的差异是否有统计学意义。

$$\chi^2_v = -2\log L_q - (-2\log L_{q+v}) = 235.5 - 219.1 = 16.4$$

由 ods 输出结果的第二部分可知, $q=2$, $q+v=10$, 所以 $v=8$ 。而 $\chi^2_{0.05(8)} = 15.51 < 16.4$, 故 $P < 0.05$ 。可认为不适合用 AR(1) 模型取代 UN 模型, 所以最后的结论应按 UN 协方差结构模型计算出来的结果来下。其假设检验结果为:

Type 3 Tests of Fixed Effects

| Effect | Num DF | Den DF | F Value | Pr > F |
|-----------------------|--------|--------|---------|--------|
| style | 1 | 16 | 14.09 | 0.0017 |
| forage | 1 | 16 | 0.73 | 0.4071 |
| style * forage | 1 | 16 | 0.14 | 0.7099 |
| time | 3 | 16 | 515.80 | <.0001 |
| style * time | 3 | 16 | 54.27 | <.0001 |
| forage * time | 3 | 16 | 17.02 | <.0001 |
| style * forage * time | 3 | 16 | 1.59 | 0.2300 |

由上述结果可知: 摄食方式(style)、观测时间(time)、摄食方式与观测时间的交互作用(style * time)及饲料种类与观测时间的交互作用(forage * time)均有统计学意义。

欲得出各均数之间两两比较的结果, 可将原程序中所有过程步删除, 换上以下过程步:

```
ods html;
proc mixed data=sastjfx10_14;
  class style forage subject time;
  model y=style|forage|time/ddfm=sat;
  repeated/type=UN sub=subject(style forage);
  lsmeans style*forage*time/pdiff;
run;
ods html close;
```

此步运行结果大部分与上述 TYPE = UN 计算结果相同, 仅多了多个均数两两比较的结果, 输出结果所占篇幅较多, 此处从略。

其实, 对于含分组因素的重复测量设计定量资料, 如果观测了各受试对象某指标的“基础”状态(即实验前状态), 对资料进行分析时, 最合适的做法是以“基础值”为定量的影响因素, 即协变量, 采用相应设计类型定量资料的协方差分析。如本例, 若第一个时间点观测的数据不是实验后第一周小鼠的体重, 而是实验前小鼠的体重, 则本资料最合适的分析方法是以实验前小鼠的体重为基础值(即协变量), 采用具有一个重复测量因素的三因素设计定量资料一元协方差分析。有关做法可参阅本章第 10.12 节。

以下是对例 10-15 的分析结果，即程序 sastjfx10_15.sas 的输出结果。

| Description | VC | CS | UN | AR1 |
|--------------------------|--------|--------|--------|--------|
| -2 Res Log Likelihood | -119.3 | -119.8 | -145.3 | -119.4 |
| AIC(smaller is better) | -117.3 | -115.8 | -103.3 | -115.4 |
| AICC(smaller is better) | -117.3 | -115.5 | -74.4 | -115.2 |
| BIC(smaller is better) | -117.0 | -115.2 | -96.9 | -114.8 |

| Description | VC | CS | UN | AR1 |
|-----------------------|----|----|----|-----|
| Covariance Parameters | 1 | 2 | 21 | 2 |
| Columns in X | 12 | 12 | 12 | 12 |
| Columns in Z | 0 | 0 | 0 | 0 |
| Subjects | 10 | 10 | 10 | 10 |
| Max Obs Per Subject | 6 | 6 | 6 | 6 |

这是上述程序中 ods 输出的结果。首先给出了 4 种协方差结构模型拟合本资料的有关情况，然后给出了协方差阵的有关信息。比较 4 种模型拟合资料情况的 AIC、BIC 数值(越小越好)，可发现 VC 协方差阵模型拟合资料较好且协方差结构中参数的个数最少。所以，最后的结论应按 VC 协方差结构模型计算出来的结果来下。其假设检验结果为：

| Type 3 Tests of Fixed Effects | | | | |
|-------------------------------|--------|--------|---------|--------|
| Effect | Num DF | Den DF | F Value | Pr > F |
| treat | 2 | 18 | 45.75 | <.0001 |
| type | 1 | 9 | 18.51 | 0.0020 |
| treat * type | 2 | 18 | 1.38 | 0.2777 |

由上述结果可知：处理(treat)和神经节类型(type)均有统计学意义。

欲得出各均数之间两两比较的结果，可将原程序中所有过程步删除，换上以下过程步：

```
ods html;
proc mixed data=sastjfx10_15;
  class patient treat type;
  model y=treat|type/ddfm=sat;
  repeated/type=VC sub=patient;
  lsmeans treat*type/pdiff;
run;
ods html close;
```

此步运行结果大部分与上述 type = VC 计算结果基本相同，仅多了多个均数两两比较的结果，输出结果所占篇幅较多，此处从略。

10.12 常见多因素实验设计一元定量资料协方差分析

10.12.1 问题与数据

【例 10-16】 某研究者欲研究 3 种饲料对动物体重增长的影响，按照某些重要非实验因素将 36 只大白鼠均分成 12 个配伍组，再将每个配伍组中的 3 只大白鼠随机分入 3 个饲料组，各组进食量与所增体重的实验结果见表 10-19。试分析三种饲料对大鼠增重效果间的差别是否有统计学意义。

表 10-19 三组白鼠的进食量 X(g) 与所增体重 Y(g) 的实验结果

| 配伍组 | 进食量(g)与所增体重(g) | | | | | |
|-----|----------------|-------|------|-------|-------|------|
| | 饲料 1: | X | Y | 饲料 2: | X | Y |
| 1 | | 256.9 | 27.0 | | 260.3 | 32.0 |
| 2 | | 271.6 | 41.7 | | 271.1 | 47.1 |
| ... | | ... | ... | | ... | ... |
| 12 | | 198.2 | 9.2 | | 199.2 | 8.1 |

【例 10-17】 某研究者将 60 只雄鼠随机分成 6 组，分别饲以不同种类食物及成分的蛋白质，并记录食物消耗量(g)、增重体重(g)，实验结果见表 10-20。试分析不同种类食物及蛋白质成分对雄鼠的增重效果之间的差异有无统计学意义。

表 10-20 6 组雄鼠的食物消耗量 X 和所增体重 Y 的实验结果

| X | Y | X | Y | X | Y | X | Y | X | Y | X | Y |
|---------|-----|-----|-----|-----|-----|--------|-----|-----|-----|-----|-----|
| 高蛋白(牛肉) | | 谷类 | | 猪肉) | | 低蛋白(牛肉 | | 谷类 | | 猪肉) | |
| 108 | 73 | 99 | 98 | 194 | 94 | 165 | 90 | 124 | 107 | 140 | 49 |
| 136 | 102 | 117 | 74 | 198 | 79 | 164 | 76 | 95 | 95 | 177 | 82 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 176 | 111 | 122 | 92 | 228 | 105 | 142 | 78 | 126 | 58 | 160 | 82 |

【例 10-18】 某研究者欲研究两种麻醉药物的效果，将 20 例病人随机均分为两组：一组用硫贲妥钠，另一组用异丙酚。分别记录 20 例病人用药前、气管插管后 1、3、5、7、9min 的收缩压变化，实验结果见表 10-21。试分析采用这两种药物麻醉的两组患者收缩压均值的差异有无统计学意义。

表 10-21 两组患者收缩压的实验结果

| 患者编号 | 药物种类 | 收缩压(mmHg) | | | | | | |
|------|------|-----------|-----|-----|-----|-----|-----|-----|
| | | 时间(min): | 0 | 1 | 3 | 5 | 7 | 9 |
| 1 | 硫贲妥钠 | | 135 | 125 | 124 | 119 | 117 | 116 |
| 2 | 硫贲妥钠 | | 116 | 127 | 111 | 104 | 104 | 99 |
| ... | ... | | ... | ... | ... | ... | ... | ... |
| 10 | 硫贲妥钠 | | 147 | 196 | 176 | 157 | 141 | 120 |
| 11 | 异丙酚 | | 124 | 100 | 98 | 98 | 90 | 84 |
| 12 | 异丙酚 | | 114 | 102 | 98 | 95 | 92 | 91 |
| ... | ... | | ... | ... | ... | ... | ... | ... |
| 20 | 异丙酚 | | 138 | 190 | 180 | 159 | 129 | 117 |

10.12.2 对数据结构的分析

例 10-16 资料中，研究者先根据某些重要非实验因素将大白鼠分成 12 个配伍组，然后再随机决定每个组中的 3 只大白鼠分别食用 3 种饲料之一。实验因素为“饲料种类”，区组因素为“某些重要非实验因素组合”，观测指标为“体重增加量”，因而资料类型应为随机区组设计一元定量资料。但在分析时，需注意“进食量”的影响，此变量为协变量。

例 10-17 资料中，涉及两个实验因素，食物种类及蛋白质成分。前者有 3 个水平：牛肉、谷类及猪肉；后者有两个水平：高蛋白和低蛋白。所有实验条件为这两个因素各水平的全面组

合且因素间无主次之分，所以资料应为两因素析因设计定量资料。但是，由于食物消耗量这个定量影响因素的存在，它是一个极为重要的非实验因素，应以其为协变量。

例 10-18 资料中，对每一个病人来说，在气管插管后 5 个时间点上分别测量其收缩压，说明“时间”因素是一个重复测量的因素。此外，还有一个实验因素“药物种类”（硫贲妥钠或异丙酚），因而这是具有一个重复测量的两因素设计定量资料，应选用具有一个重复测量的两因素设计定量资料方差分析来处理。同时，由于研究者记录了所有病人麻醉前的收缩压值，所以最好以此为“基础值”或协变量的取值。

10.12.3 分析目的与统计分析方法的选择

对于例 10-16 资料，资料类型为随机区组设计一元定量资料，但由于在分析时还要考虑定量影响因素“进食量”的影响，需选用随机区组设计定量资料一元协方差分析。

对于例 10-17 资料，资料类型为两因素析因设计一元定量资料。但是，由于食物消耗量这个定量影响因素的存在，分析时应以食物消耗量为协变量，采用两因素析因设计一元定量资料的协方差分析处理此资料。

对于例 10-18 资料，资料类型为具有一个重复测量的两因素设计一元定量资料。由于研究者记录了所有病人麻醉前的收缩压值，所以最好以此为“基础值”，采用具有一个重复测量的两因素设计一元定量资料的协方差分析来处理数据。

10.12.4 SAS 程序

对例 10-16 资料进行随机区组设计定量资料一元协方差分析，SAS 程序名为 sastjfx10_16.sas。

```
data sastjfx10_16;          /* 1 */
  do group = 1 to 12;
    do forage = 1 to 3;
      input appetite increment @@ ;
      output;
    end; end;
  cards;
256.9 27.0 260.3 32.0 544.7 160.3
271.6 41.7 271.1 47.1 481.2 96.1
. . . . .
198.2 9.2 199.2 8.1 371.9 54.3
;
run;
ods html;
```

```
proc glm;                  /* 2 */
  class group forage;
  model increment = appetite group
    forage appetite* group
    appetite* forage/ss3;
run;

proc glm;                  /* 3 */
  class group forage;
  model increment = appetite group
    forage/solution ss3;
  lsmeans forage/stderr tdiff pdiff;
run;
ods html close;
```

程序第 1 步建立数据集，group 代表“配伍组”，forage 代表“饲料种类”，appetite 代表“进食量”，increment 代表“体重增加量”。第 2 步调用 glm 过程分析协变量与实验因素之间的交互作用是否有统计学意义，目的是了解各组的总体回归斜率是否相等。第 3 步调用 glm 过程进行随机区组设计定量资料的协方差分析。model 语句“/”后的 solution 选项用来给出模型中固定效应的解，lsmeans 语句可给出响应变量 increment 的修正均数，并给出因素各水平两两比较的结果。lsmeans 语句“/”后的 stderr 用来输出因素各水平组修正后响应变量的标准误及修正均数与 0 比较的检验结果，tdiff 和 pdiff 用来输出因素各水平组修正均数两两比较的 t 值和 P 值。

对例 10-17 资料进行两因素析因设计一元定量资料的协方差分析, SAS 程序名为 sastjfx10_17.sas。

```

data sastjfx10_17;          /* 1 */
  do protein=1 to 2;
  do food=1 to 3;
  input appetite increment@@ ;
  output;
  end;
  end;
  cards;
108 73 99 98 194 94 165 90
124 107 140 49 136 102 117 74
. . . . .
228 105 142 78 126 58 160 82
;
run;
ods html;

proc glm;                  /* 2 */
  class protein food;
  model increment = appetite protein
    food protein*food
    appetite* protein appetite*
    food/ ss3;
run;

proc glm;                  /* 3 */
  class protein food;
  model increment = appetite
    protein food protein* food/ss3;
  lsmeans protein food protein* food
  /stderr tdiff pdiff;
run;
ods html close;

```

程序第 1 步建立数据集, protein 代表“蛋白质成分”, food 代表“食物种类”, appetite 代表“食物消耗量”, increment 代表“增重体重”。第 2 步调用 glm 过程分析协变量与两个实验因素之间的交互作用是否有统计学意义, 目的是了解各组的总体回归斜率是否相等。第 3 步调用 glm 过程进行析因设计定量资料的一元协方差分析。

对例 10-18 资料进行具有一个重复测量的两因素设计一元定量资料的协方差分析, SAS 程序名为 sastjfx10_18.sas。

```

data sastjfx10_18;          /* 1 */
  do drug=1 to 2;
  do patient=1 to 10;
  input time0@@ ;
  do time=1,3,5,7,9;
  input y @@ ;
  output;
  end; end; end;
  cards;
135 125 124 119 117 116
116 127 111 104 104 99
110 142 128 122 105 98
135 148 139 128 117 110
125 152 128 109 101 100
128 155 128 115 104 100
136 159 137 123 96 94
138 168 164 144 120 116
142 182 156 128 127 111
147 196 176 157 141 120
124 100 98 98 90 84
114 102 98 95 92 91
130 111 113 110 108 99
128 115 114 107 101 102
134 121 118 100 96 97
107 122 109 105 98 101

proc mixed data =sastjfx10_18; /* 4 */
  class time0 drug patient time;
  model y=time0 drug |time;
  repeated/type=CS
  sub=patient(drug);
  ods output fitstatistics=b;
  ods output dimensions=b1;
run;

proc mixed data =sastjfx10_18; /* 5 */
  class time0 drug patient time;
  model y=time0 drug |time;
  repeated/type=AR(1)
  sub=patient(drug);
  ods output fitstatistics=c;
  ods output dimensions=c1;
run;

proc mixed data =sastjfx10_18; /* 6 */
  class time0 drug patient time;
  model y=time0 drug |time;
  repeated/type=SP(POW)(time)
  sub=patient(drug);
  ods output fitstatistics=d;

```

```

120 142 109 96 93 90
125 143 129 119 116 113
126 164 122 110 100 92
138 190 180 159 129 117
run;
ods html;
proc glm; /* 2 */
  class drug time;
  model y=time0 drug|time
        time0* drug time0* time/ss3;
run;
ods html close;
proc mixed data=sastjfx10_18; /* 3 */
  class time0 drug patient time;
  model y=time0 drug|time;
  repeated/type = VC sub = patient
(drug);
  ods output fitstatistics=a;
  ods output dimensions=a1;
run;

ods output dimensions=d1;
run;
%macro shuju(dataset,y); /* 7 */
  data &dataset;
  set &dataset;
  rename value = &y;
run;
%mend shuju;
%shuju(a,VC) %shuju(b,CS)
%shuju(c,AR1) %shuju(d,SP)
%shuju(a1,VC) %shuju(b1,CS)
%shuju(c1,AR1) %shuju(d1,SP)
data f; /* 8 */
  merge a b c d;
run;
data f1; /* 9 */
  merge a1 b1 c1 d1;
run;
ods html;
proc print data=f; /* 10 */
  format_numeric_5.1;
run;
proc print data=f1; /* 11 */
run;
ods html close;

```

程序第1步建立数据集，drug代表“药物种类”，patient代表“患者编号”，time0代表患者用药前收缩压值，time代表气管插管后时间，y代表收缩压值。第2步调用glm过程分析协变量与两个实验因素之间的交互作用是否有统计学意义，目的是了解各组的总体回归斜率是否相等。第3、4、5、6步调用mixed过程，分别采用VC、CS、AR(1)、SP(POW)四种协方差结构模型对资料进行方差分析(此数据不宜采用UN协方差结构模型进行方差分析，因其迭代无法收敛，读者可自行验证)。第7步建立宏shuju，以实现数据集中已有变量value的更名，具体语法读者可参考第3章第3.2节有关内容。第8、9步均用来实现对不同数据集的横向合并。第10、11步均用来将数据集中的内容输出到output窗口中去。

10.12.5 主要分析结果及解释

以下是对例10-16的分析结果，即程序sastjfx10_16.sas的输出结果。

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|-------------------|----|-------------|-------------|---------|--------|
| appetite | 1 | 31.802446 | 31.802446 | 0.48 | 0.5073 |
| group | 11 | 1456.754538 | 132.432231 | 2.01 | 0.1662 |
| forage | 2 | 14.989352 | 7.494676 | 0.11 | 0.8941 |
| appetite * group | 11 | 1631.176555 | 148.288778 | 2.25 | 0.1301 |
| appetite * forage | 2 | 45.164419 | 22.582210 | 0.34 | 0.7202 |

这是输出结果的第1部分，用来考察资料是否满足协方差分析的第2个前提条件——各组总体回归斜率相等。查看上述结果可发现：appetite * forage对应的假设检验结果为 $F=0.34$ ， $P=0.7202$ ；appetite * group对应的假设检验结果为 $F=2.25$ ， $P=0.1301$ 。appetite * forage和appetite

* group 均无统计学意义, 可认为各回归直线之间的斜率相等, 所以满足协方差分析的第 2 个前提条件。对于此方差分析表中其他各项主效应的假设检验结果, 读者可不予参考。

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|----------|----|-------------|-------------|---------|--------|
| appetite | 1 | 6174.248301 | 6174.248301 | 58.26 | <.0001 |
| group | 11 | 3765.325658 | 342.302333 | 3.23 | 0.0101 |
| forage | 2 | 463.947755 | 231.973878 | 2.19 | 0.1369 |

这是输出结果的第 2 部分, 是对 3 个因素进行假设检验的结果。由各自对应的 F 值和相应的 P 值大小可判断, 进食量 appetite ($F = 58.26$, $P < .0001$)、各配伍组之间 ($F = 3.23$, $P = 0.0101$) 对响应变量 increment 的影响有统计学意义, 3 个饲料组 ($F = 2.19$, $P = 0.1369$) 观测指标 increment 均数之间的差别无统计学意义。

| Parameter | Estimate | | Standard Error | t Value | Pr > t |
|-----------|--------------|---|----------------|---------|---------|
| Intercept | -89.08418366 | B | 22.48817353 | -3.96 | 0.0007 |
| appetite | 0.40881283 | | 0.05355788 | 7.63 | <.0001 |
| group 1 | 9.36045515 | B | 9.89612611 | 0.95 | 0.3550 |
| group 2 | 3.07208435 | B | 9.55542006 | 0.32 | 0.7510 |
| group 3 | 24.74781468 | B | 8.50972658 | 2.91 | 0.0084 |
| group 4 | 7.26167590 | B | 10.88594221 | 0.67 | 0.5120 |
| group 5 | -2.00816529 | B | 8.86073025 | -0.23 | 0.8229 |
| group 6 | -7.38413810 | B | 9.68211904 | -0.76 | 0.4541 |
| group 7 | 15.43805935 | B | 9.11888535 | 1.69 | 0.1052 |
| group 8 | -6.07055235 | B | 9.82518315 | -0.62 | 0.5433 |
| group 9 | 10.96065907 | B | 9.91691397 | 1.11 | 0.2816 |
| group 10 | -0.54852862 | B | 12.55752149 | -0.04 | 0.9656 |
| group 11 | 23.48757974 | B | 12.30629843 | 1.91 | 0.0701 |
| group 12 | 0.00000000 | B | . | . | . |
| forage 1 | 8.36529178 | B | 12.51818204 | 0.67 | 0.5113 |
| forage 2 | 15.98754776 | B | 12.39759817 | 1.29 | 0.2112 |
| forage 3 | 0.00000000 | B | . | . | . |

这是输出结果的第 3 部分, 是模型中固定效应的解。对区组因素 group 和实验因素饲料种类 forage 来说, 均是其各自水平与参照水平 (各因素最后一个水平, 本资料为 group12 和 forage3) 进行均数比较的假设检验结果。如 forage1 所在行对应的 P 值为 0.5113, 即表示第 1 种饲料与第 3 种饲料之间的差别无统计学意义。在 model 语句中不存在交互项时, 这个结果与随后 lsmeans 语句给出的结果相同, 可视为后者的一部分。所以, 如果 SAS 程序中已有 lsmeans 语句, 即可省去 model 语句“/”后的 solution 项。

Least Squares Means

| forage | increment LSMEAN | Standard Error | Pr > t | LSMEAN Number |
|--------|------------------|----------------|---------|---------------|
| 1 | 67.4282342 | 4.9616061 | <.0001 | 1 |
| 2 | 75.0504901 | 4.8596353 | <.0001 | 2 |
| 3 | 59.0629424 | 8.3641102 | <.0001 | 3 |

Least Squares Means for effect forage

Pr > |t| for $H_0: \text{LSMean}(i) = \text{LSMean}(j)$

| Dependent Variable: increment | | | |
|-------------------------------|----------|----------|----------|
| i/j | 1 | 2 | 3 |
| 1 | | -1.81287 | 0.668251 |
| | | 0.0842 | 0.5113 |
| 2 | 1.81287 | | 1.289568 |
| | 0.0842 | | 0.2112 |
| 3 | -0.66825 | -1.28957 | |
| | 0.5113 | 0.2112 | |

这是输出结果的第4部分,首先给出了3个饲料组响应变量 increment 的修正均数及其与0比较的假设检验的结果,无太大实际意义。然后给出3个饲料组响应变量 increment 的修正均数两两比较的结果,3种饲料间的差别均无统计学意义。

以下是对例10-17的分析结果,即程序 sastjfx10_17.sas 的输出结果。

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|--------------------|----|-------------|-------------|---------|--------|
| appetite | 1 | 2277.195663 | 2277.195663 | 13.45 | 0.0006 |
| protein | 1 | 28.318754 | 28.318754 | 0.17 | 0.6843 |
| food | 2 | 155.117871 | 77.558935 | 0.46 | 0.6351 |
| protein * food | 2 | 857.568437 | 428.784219 | 2.53 | 0.0896 |
| appetite * protein | 1 | 112.418396 | 112.418396 | 0.66 | 0.4190 |
| appetite * food | 2 | 10.890008 | 5.445004 | 0.03 | 0.9684 |

这是输出结果的第1部分,用来考察资料是否满足协方差分析的第2个前提条件——各组总体回归斜率相等。查看上述结果可发现,appetite * protein 对应的假设检验结果为 $F = 0.66$ 、 $P = 0.4190$, appetite * food 对应的假设检验结果为 $F = 0.03$ 、 $P = 0.9684$,即 appetite * forage 和 appetite * food 无统计学意义。可认为蛋白质成分及食物种类这2个定性变量内部的回归斜率近似相等,所以满足协方差分析的第2个前提条件。对于此方差分析表中其他各项的假设检验结果,读者可不予参考。

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|----------------|----|-------------|-------------|---------|--------|
| appetite | 1 | 2990.626114 | 2990.626114 | 18.44 | <.0001 |
| protein | 1 | 2343.462515 | 2343.462515 | 14.45 | 0.0004 |
| food | 2 | 1673.305084 | 836.652542 | 5.16 | 0.0090 |
| protein * food | 2 | 933.811700 | 466.905850 | 2.88 | 0.0650 |

这是输出结果的第2部分,是对协变量、2个实验因素及其交互作用进行假设检验的结果。由各自对应的F值和相应的P值大小可判断,appetite、protein及food对响应变量 increment 的影响均有统计学意义,但 protein * food 无统计学意义。

| Least Squares Means | | | | | |
|---------------------|------------------|----------------|---------------|----------------------|---------|
| protein | increment | Standard | H0:LSMEAN = 0 | H0:LSMean1 = LSMean2 | |
| | LSMEAN | Error | Pr > t | t Value | Pr > t |
| 1 | 94.1737464 | 2.3357717 | <.0001 | 3.80 | 0.0004 |
| 2 | 81.5595869 | 2.3357717 | <.0001 | | |
| food | increment LSMEAN | Standard Error | Pr > t | LSMEAN Number | |
| 1 | 89.9826573 | 2.8489970 | <.0001 | 1 | |

| | | | | |
|---|------------|-----------|--------|---|
| 2 | 98.7404219 | 4.3007776 | <.0001 | 2 |
| 3 | 74.8769208 | 4.3679550 | <.0001 | 3 |

Least Squares Means for effect food

Pr > |t| for H0: LSMean(i) = LSMean(j)

Dependent Variable: increment

| i/j | 1 | 2 | 3 |
|-----|--------------------|--------------------|--------------------|
| 1 | | -1.71625 0.0920 | 2.865683 0.0060 |
| 2 | 1.71625 0.0920 | | 3.108718 0.0030 |
| 3 | -2.86568 0.0060 | -3.10872 0.0030 | |

这是输出结果的第 3 部分, 包括 protein、food 两因素各水平下响应变量 increment 的修正均数、与 0 比较以及修正均数之间两两比较的假设检验结果。除 food 因素 1 水平和 2 水平(即牛肉与谷类)条件下响应变量 increment 的修正均数之间的差别无统计学意义外, 其他各水平条件下响应变量 increment 的修正均数之间的差别均有统计学意义。

| protein | food | increment LSMEAN | Standard Error | Pr > t | LSMEAN Number |
|---------|------|------------------|----------------|---------|---------------|
| 1 | 1 | 101.548290 | 4.043227 | <.0001 | 1 |
| 1 | 2 | 100.764766 | 5.310373 | <.0001 | 2 |
| 1 | 3 | 80.208183 | 6.033249 | <.0001 | 3 |
| 2 | 1 | 78.417024 | 4.031244 | <.0001 | 4 |
| 2 | 2 | 96.716078 | 5.012465 | <.0001 | 5 |
| 2 | 3 | 69.545659 | 4.556548 | <.0001 | 6 |

Least Squares Means for effect protein * food

Pr > |t| for H0: LSMean(i) = LSMean(j)

Dependent Variable: increment

| i/j | 1 | 2 | 3 | 4 | 5 | 6 |
|-----|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| 1 | 0.120826 0.9043 | 2.852015 0.0062 | 4.043205 0.0002 | 0.770615 0.4444 | 5.14787 <.0001 | |
| 2 | -0.12083 0.9043 | | 2.10131 0.0404 | 3.305311 0.0017 | 0.708413 0.4818 | 3.910932 0.0003 |
| 3 | -2.85201 0.0062 | -2.10131 0.0404 | | 0.250781 0.8030 | -1.75635 0.0848 | 1.7295 0.0895 |
| 4 | -4.04321 0.0002 | -3.30531 0.0017 | -0.25078 0.8030 | | -2.80813 0.0070 | 1.473745 0.1465 |
| 5 | -0.77061 0.4444 | -0.70841 0.4818 | 1.756354 0.0848 | 2.808134 0.0070 | | 3.548995 0.0008 |
| 6 | -5.14787 <.0001 | -3.91093 0.0003 | -1.7295 0.0895 | -1.47374 0.1465 | -3.54899 0.0008 | |

这是输出结果的第 4 部分, 首先给出了各实验条件下响应变量 increment 的修正均数及它们与 0 之间的差异是否有统计学意义的假设检验结果, 并对每种实验条件进行编号(查看

LSMEAN Number 列), 然后给出了 6 种实验条件下响应变量 increment 的修正均数两两比较的结果。编号为 1~3 的实验条件之间的比较是高蛋白条件下 3 种食物营养价值的比较: 1(牛肉)与 2(谷类)之间的差别无统计学意义($P=0.9043$), 1 与 3、2 与 3 之间的差别均有统计学意义。编号为 4~6 的实验条件之间的比较是低蛋白条件下 3 种食物营养价值的比较: 4(牛肉)与 6(猪肉)之间的差别无统计学意义($P=0.1465$), 4 与 5、5 与 6 之间的差别均有统计学意义。说明高蛋白时, 牛肉与谷类的营养价值接近且最高; 低蛋白时, 谷类的营养价值最高, 牛肉与猪肉营养价值接近且最低。

以下是对例 10-18 的分析结果, 即程序 sastjfx10_18.sas 的输出结果。

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|--------------|----|-------------|-------------|---------|--------|
| time0 | 1 | 10030.34009 | 10030.34009 | 44.01 | <.0001 |
| drug | 1 | 28.39311 | 28.39311 | 0.12 | 0.7250 |
| time | 4 | 616.91155 | 154.22789 | 0.68 | 0.6099 |
| drug * time | 4 | 344.43027 | 86.10757 | 0.38 | 0.8239 |
| time0 * drug | 1 | 4.07882 | 4.07882 | 0.02 | 0.8939 |
| time0 * time | 4 | 1203.37074 | 300.84268 | 1.32 | 0.2692 |

这是输出结果的第 1 部分, 用来考察资料是否满足协方差分析的第 2 个前提条件——各组总体回归斜率相等。查看上述结果可发现, time0 * drug 对应的假设检验结果为 $F=0.02$, $P=0.8939$, time0 * time 对应的假设检验结果为 $F=1.30$, $P=0.2692$ 。time0 * drug 和 time0 * time 无统计学意义, 可认为药物种类及时间这两个定性变量内部的回归斜率近似相等, 所以满足协方差分析的第 2 个前提条件。对于此方差分析表中其他各项的假设检验结果, 读者可不予参考。

| Obs | Descr | VC | CS | AR1 | SP |
|-----|-------------------------|-------|-------|-------|-------|
| 1 | -2 Res Log Likelihood | 611.4 | 608.3 | 565.8 | 565.8 |
| 2 | AIC(smaller is better) | 613.4 | 612.3 | 569.8 | 569.8 |
| 3 | AICC(smaller is better) | 613.5 | 612.5 | 570.0 | 570.0 |
| 4 | BIC(smaller is better) | 614.4 | 614.3 | 571.8 | 571.8 |

| Obs | Descr | VC | CS | AR | SP |
|-----|-----------------------|----|----|----|----|
| 1 | Covariance Parameters | 1 | 2 | 2 | 2 |
| 2 | Columns in X | 34 | 34 | 34 | 34 |
| 3 | Columns in Z | 0 | 0 | 0 | 0 |
| 4 | Subjects | 20 | 20 | 20 | 20 |
| 5 | Max Obs Per Subject | 5 | 5 | 5 | 5 |

这是上述程序中 ods(output delivery system)输出的结果。首先给出了 4 种协方差结构模型拟合本资料的有关情况, 然后给出了协方差阵的有关信息(Covariance Parameters 表示模型中待估计的协方差结构中参数的个数)。比较 4 种模型拟合资料情况的 AIC、BIC 数值, 可发现 AR(1)和 SP(POW)两种协方差阵模型拟合资料情况相同且较好。由于 CS、AR(1)、SP(POW)三种协方差阵模型参数个数均为 2, 但后两者拟合效果好于 CS 协方差阵模型, 所以可不考虑 CS 协方差阵模型。现比较 AR(1)和 SP(POW)两种协方差阵模型拟合资料的效果与 VC 协方差阵模型拟合资料的效果之间的差异是否有统计学意义。

$$\chi^2_v = -2\log L_q - (-2\log L_{q+v}) = 611.4 - 565.8 = 45.6$$

由 ods 输出结果的第 2 部分可知, $q = 1$, $q + v = 2$, 所以 $v = 1$ 。而 $\chi^2_{0.05(1)} = 3.84 < 45.6$, 故 $P < 0.05$ 。可认为不适合用 VC 模型取代 AR(1) 或 SP(POW) 模型, 所以最后的结论应按 AR(1) 或 SP(POW) 协方差结构模型计算出来的结果来下。其假设检验结果为:

| Type 3 Tests of Fixed Effects | | | | |
|-------------------------------|--------|--------|---------|---------|
| Effect | Num DF | Den DF | F Value | Pr > F |
| time0 | 15 | 3 | 0.81 | 0.6686 |
| drug | 1 | 3 | 0.00 | 0.9672 |
| time | 4 | 72 | 33.04 | < .0001 |
| drug * time | 4 | 72 | 1.30 | 0.2794 |

由上述结果可知: 时间(time)因素各水平之间观测指标的差异有统计学意义, 而药物种类(drug)、药物种类与时间的交互作用(drug * time)均无统计学意义。所以, 两种药物麻醉后患者的收缩压状况没有差异。

10.13 多个单因素 2 水平设计定量资料 meta 分析

10.13.1 问题与数据

【例 10-19】 为了应用 meta 分析确定国内成人哮喘患者血清总 IgE 含量的平均范围, 以便临床哮喘诊断、分型及预后估计时参考, 宋士德等人通过计算机光盘检索哮喘研究文献 2000 余篇, 其中有关成人哮喘患者血清总 IgE 含量测定研究的文献占 102 篇。再从中按照此 meta 分析的设计要求(纳入 meta 分析的条件略)进行筛选。经筛选, 得到 21 项符合要求的、来自国内多种期刊、分布于全国各地的研究结果(如表 10-22 所示)。请对该资料进行 meta 分析。

表 10-22 21 项国内成年哮喘患者血清总 IgE 水平(IU/mL)的研究结果

| 研究编号 | 实验组 | | | 对照组 | | |
|------|-----|-----------|-----|-----|-----------|-----|
| | n | \bar{x} | s | n | \bar{x} | s |
| 1 | 30 | 1075 | 507 | 30 | 150 | 130 |
| 2 | 33 | 1686 | 297 | 157 | 455 | 70 |
| ... | ... | ... | ... | ... | ... | ... |
| 21 | 26 | 743 | 343 | 26 | 230 | 89 |

资料来源: 宋士德等. 国人哮喘患者血清 IgE 水平的 meta 分析. 数理医药学杂志, 1998, 11(3): 232-234.

10.13.2 对数据结构的分析

研究者共收集了 21 项符合纳入标准的研究, 对每一个项目而言, 其设计类型都是单因素 2 水平设计。

10.13.3 分析目的与统计分析方法的选择

研究者希望通过对这 21 项研究及其结果进行综合分析和再评价, 从而得出国内成人哮喘患者血清总 IgE 含量的平均范围。合适的统计分析方法是多个单因素 2 水平设计定量资料的 meta 分析。

10.13.4 SAS 程序

对本资料进行单因素 2 水平设计一元定量资料 meta 分析, SAS 程序名为 sastjfx10_19.sas。

```

DATA SASTJFX10_19;
INPUT n_Ti xbar_Ti s_Ti n_Ci xbar_Ci
s_Ci @@ ;
Ni = n_Ti + n_Ci;
s_combined = SQRT(((n_Ti - 1) * s_Ti** 2
+ (n_Ci - 1) * s_Ci** 2) /
(n_Ti + n_Ci - 2));
di = ((xbar_Ti - xbar_Ci) /
s_combined)
* (1 - 3 / (4 * Ni - 9));
wi = 2 * Ni / (8 + di** 2);
widi = wi * di; widi2 = wi * di** 2;
number_of_study = 21;
CARDS;
30 1075 507 30 150 130
33 1686 297 157 455 70
. . . . .
26 743 343 26 230 89
;
RUN;
DATA A; SET SASTJFX10_19;
Sum_of_Ni_temp + Ni;
Sum_of_wi_temp + wi;
Sum_of_wi2_temp + wi * wi;
Sum_of_widi_temp + widi;
Sum_of_widi2_temp + widi2;
ID = _N_;
IF _N_ = number_of_study THEN DO;
Sum_of_Ni = Sum_of_Ni_temp;
Sum_of_wi = Sum_of_wi_temp;
Sum_of_wi2 = Sum_of_wi2_temp;
Sum_of_widi = Sum_of_widi_temp;
Sum_of_widi2 = Sum_of_widi2_temp;
w_bar = Sum_of_wi / number_of_study;
Square_of_s_w = (Sum_of_wi2 - number_of_study * w_bar** 2) / (number_of_study - 1);
U = (number_of_study - 1) * (w_bar - Square_of_s_w / (number_of_study * w_bar));
Q = Sum_of_widi2 - (Sum_of_widi)** 2 / Sum_of_wi;
Square_of_tao_hat = (Q - (number_of_study - 1)) / U;

IF Square_of_tao_hat < 0
THEN Square_of_tao_hat = 0;
d_combined = Sum_of_widi / Sum_of_wi;
DF = number_of_study - 1;
P = 1 - PROBCHI(Q, DF);
I2 = (Q - (number_of_study - 1)) / Q;
IF I2 < 0 THEN I2 = 0;
FLOW = d_combined - 1.96 / Sum_of_wi** 0.5;
FUP = d_combined + 1.96 / Sum_of_wi** 0.5;
END; RUN;
PROC SORT; BY DESCENDING ID;
DATA B; SET A;
Square_of_tao_hat_for_all
+ Square_of_tao_hat;
wi_bar = 1 / (1 / wi + Square_of_tao_hat_for_all);
RUN;
PROC SORT; BY ID;
DATA C; SET B;
Sum_of_wi_bar_temp + wi_bar;
Sum_of_wi_bardi_temp + wi_bar * di;
IF _N_ = number_of_study THEN DO;
Sum_of_wi_bar = Sum_of_wi_bar_temp;
Sum_of_wi_bardi = Sum_of_wi_bardi_temp;
V_hat_d_bar_hat = 1 / Sum_of_wi_bar;
d_bar_hat = Sum_of_wi_bardi / Sum_of_wi_bar;
LOW = d_bar_hat - 1.96 *
V_hat_d_bar_hat** 0.5;
UP = d_bar_hat + 1.96 *
V_hat_d_bar_hat** 0.5;
END; RUN;
PROC SORT; BY DESCENDING ID;
DATA D; SET C;
WHERE ID = number_of_study; RUN;
PROC PRINT;
VAR number_of_study Q DF P I2
d_combined FLOW FUP d_bar_hat
LOW UP;
ODS HTML; RUN;

```

程序中分别用 n_{Ti} 、 \bar{x}_{Ti} 和 s_{Ti} 表示各研究中实验组的样本含量、样本均值和样本标准差；分别用 n_{Ci} 、 \bar{x}_{Ci} 和 s_{Ci} 表示各研究中对照组的样本含量、样本均值和样本标准差；用 $s_{combined}$ 表示各研究的混合标准差；用 d_i 表示各研究的效应（此处用标准均差表示）；用 N_i 表示各研究的样本含量；用 w_i 表示采用固定效应模型时的权重；用 $widi$ 表示 w_i 与 d_i 的乘积；用 $widi2$ 表示 w_i 与 d_i 的平方的乘积；用 $number_of_study$ 表示纳入研究的数量；用 $Sum_of_Ni_temp$ 表示 N_i 累积的中间结果，用 Sum_of_Ni 表示 N_i 累积的最终结果；用 $Sum_of_wi_temp$ 、 $Sum_of_wi2_temp$ 、 $Sum_of_widi_temp$ 和 $Sum_of_widi2_temp$ 分别表示 w_i 、 $wi2$ 、 $widi$ 和 $widi2$ 累

积的中间结果, 用 Sum_of_wi、Sum_of_wi2、Sum_of_widi 和 Sum_of_widi2 分别表示 wi、wi2、widi 和 widi2 的累积的最终结果; 用 d_combined 和 d_bar_hat 分别表示采用固定效应模型和随机效应模型时的合并效应; 用 w_bar、Square_of_s_w、Square_of_tao_hat 和 wi_bar 分别表示 \bar{w} 、 s_w^2 、 τ^2 和 \bar{w}_i ; 用 DF 表示自由度; 用 Q 和 P 分别表示异质性检验的 χ^2 值和 P 值; I2 为异质性检验的另一个统计量; 用 Sum_of_wi_bar_temp 和 Sum_of_wi_bardi_temp 分别表示 wi_bar 和 wi_bardi 累积的中间结果, 用 Sum_of_wi_bar 和 Sum_of_wi_bardi 分别表示 wi_bar 和 wi_bardi 累积的最终结果; 用 V_hat_d_bar_hat 表示 $\hat{V}(\hat{d})$; 用 FLOW、FUP 及 LOW、UP 分别表示采用固定效应模型和采用随机效应模型时合并效应的 95% 置信区间的下限和上限。

值得一提的是, 程序中多个变量是用下划线将几个单词连接在一起形成的。

10.13.5 主要分析结果及解释

| Obs | number_of_study | | Q | DF | P | I2 |
|------------|-----------------|---------|-----------|---------|---|---------|
| 1 | 21 | | 541.309 | 20 | 0 | 0.96305 |
| d_combined | FLOW | FUP | d_bar_hat | LOW | | UP |
| 2.58455 | 2.44612 | 2.72299 | 3.79902 | 3.05436 | | 4.54369 |

异质性检验的结果为: $Q = \chi^2 = 541.309$, $df = 21 - 1 = 20$, $PI^2 = 0.96305 > 0.5$, 因此认为各研究同质性很差, 应采用随机效应模型进行综合分析, 结果表明: 平均效应尺度(SMD)为 3.80 倍标准差, 其 95% 置信区间为 3.05 ~ 4.54 倍标准差。由于置信区间位于 0 的右侧, 所以 21 项研究的综合结果表明, 哮喘对成人血清 IgE 水平的影响有统计学意义。专业结论为: 哮喘对成人血清 IgE 水平具有严重影响, 成人患哮喘后血清 IgE 水平显著上升, 平均上升水平为 3.80 倍标准差, 其 95% 置信区间为 3.05 ~ 4.54 倍标准差。

10.14 本章小结

本章主要介绍了常见的多因素设计一元定量资料的有关概念、分析方法及 SAS 实现, 涵盖的设计类型包括随机区组设计、双因素无重复设计、平衡不完全区组设计、拉丁方设计、交叉设计、析因设计、含区组因素的析因设计、裂区设计、嵌套设计、正交设计和重复测量设计。统计分析方法主要以定量资料方差分析为主, 部分设计给出了非参数检验的方法。

另外, 还介绍了单因素 2 水平设计一元定量资料在满足可合并性条件时的 meta 分析方法及其 SAS 实现, 并给出了具体实例及相应的 SAS 程序。但须注意两点: (1) meta 分析方法本身还不完善; (2) 只有对具有一定的可合并性的多项研究资料才可做 meta 分析, 不同设计类型的资料不可以混在一起做 meta 分析。

学习本章, 可从几个方面入手: 熟练辨析实验设计类型, 对各种常见实验设计的特点和它们之间的相互区别有较为深刻的理解; 了解方差分析应用的前提条件和变异分解的有关原理; 明白实现统计分析的 SAS 程序各步的用途, 并能看懂 SAS 程序的输出结果; 结合统计分析的结果和专业知识, 得出相关的统计学结论和专业结论。

本章对随机区组设计定量资料、双因素无重复设计定量资料的统计分析给出了正态性和方差齐性检验的过程。为节省篇幅, 其余各种设计类型的定量资料均假设满足这两个条件而未加以考察。

(高辉 周诗国 张泽)

第 11 章 单因素设计多元定量资料差异性分析

在某种实验设计类型之下，若每次只分析一个或多个定性影响因素对一个定量指标的影响，常采用一元方差分析；若每次用参数法同时分析一个或多个定性影响因素对两个或两个以上在专业上有一定联系的定量指标的影响情况时，就称为多元方差分析(Multivariate Analysis of Variance, MANOVA)。做方差分析时，影响因素都是定性的。当除了定性的影响因素之外还有定量的影响因素存在(或者不可忽略)时，要分析各影响因素对定量指标的影响，则需要采用另外一种统计学分析方法，即协方差分析(Analysis of Covariance)才能达到目的。本章将介绍用 SAS 实现单因素设计多元定量资料多元方差和协方差分析方法。

11.1 问题、数据及统计分析方法的选择

11.1.1 问题与数据

【例 11-1】 用益寿宁治疗 5 名高血脂患者，治疗结果如表 11-1 所示。假定定量资料满足参数检验的前提条件。问：该药是否有降血压的作用？

【例 11-2】 某医生观测了 20 名正常人的动态心电图，分析出各受试对象早间和下午的低频(LF)及高频(HF)心电频谱值，结果见表 11-2(部分数据省略，完整数据详见本书附带光盘中的数据文件，下同)。试问两时段的频谱值差别有无统计学意义？

表 11-1 5 名高血脂患者经益寿宁治疗后
胆固醇和三酰甘油下降的情况

| 受试者 编号 | 胆固醇下降值 X_1 (mg%) | 三酰甘油下降值 X_2 (mg%) |
|-----------|-----------------------|------------------------|
| 1 | 16 | -4 |
| 2 | 21 | 46 |
| 3 | 57 | -40 |
| 4 | -20 | 107 |
| 5 | 17 | 86 |

表 11-2 20 名正常人早间和下午
的低、高频心电频谱结果

| 个体编号 | LF | | HF | |
|------|-------|-------|------|-------|
| | 早间 | 下午 | 早间 | 下午 |
| 1 | 105.7 | 115.9 | 18.0 | 64.8 |
| 2 | 93.5 | 208.9 | 14.2 | 119.4 |
| ... | ... | ... | ... | ... |
| 19 | 132.1 | 187.0 | 41.6 | 44.4 |
| 20 | 205.8 | 105.4 | 66.2 | 81.6 |

【例 11-3】 某人在某地按性别和年龄分层随机抽取 30 名初生到 3 周岁的婴幼儿，并测量了他们的身高(X_1)、体重(X_2)和体表面积(Y)，资料见表 11-3。问：在初生到 3 周岁的婴幼儿中，男孩和女孩在三项指标上总的来说差异有无统计学意义？

表 11-3 30 名婴幼儿的身高、体重及体表面积

| 性别 | 编号 | 身高(cm) | 体重(kg) | 体表面积(cm^2) |
|----|-----|--------|--------|-----------------------|
| 男 | 1 | 54.0 | 3.0 | 2446.2 |
| | 2 | 50.5 | 2.3 | 1928.4 |
| | ... | ... | ... | ... |
| | 14 | 92.0 | 11.0 | 5283.3 |
| | 15 | 94.0 | 15.0 | 6101.6 |
| 女 | 16 | 54.0 | 3.0 | 2117.3 |
| | 17 | 53.0 | 2.3 | 2200.2 |
| | ... | ... | ... | ... |
| | 29 | 92.0 | 12.0 | 5299.4 |
| | 30 | 91.0 | 12.5 | 5291.5 |

【例 11-4】 某研究者测得某地 43 名 9~11 岁小学生的提转敏捷度、目标追踪两

项神经行为功能指标的得分,如表 11-4 所示。欲比较三个不同水平上的小学生的两项指标是否相同,请做统计分析(假定资料满足参数检验的前提条件)。

表 11-4 某地三个不同成绩水平小学生的提转敏捷度得分(y_1)和目标追踪得分(y_2)

| 成绩水平 | 编号 | y_1 | y_2 | 成绩水平 | 编号 | y_1 | y_2 | 成绩水平 | 编号 | y_1 | y_2 |
|------|-----|-------|-------|------|-----|-------|-------|------|-----|-------|-------|
| 一般 | 1 | 63 | 62 | 较好 | 17 | 49 | 102 | 极好 | 33 | 73 | 110 |
| | 2 | 54 | 72 | | 18 | 49 | 94 | | 34 | 74 | 62 |
| | ... | ... | ... | | ... | ... | ... | | ... | ... | ... |
| | 15 | 55 | 110 | | 31 | 62 | 93 | | 42 | 56 | 127 |
| | 16 | 47 | 87 | | 32 | 66 | 100 | | 43 | 77 | 131 |

【例 11-5】 为了研究学生的性别(1 = 男, 2 = 女)、年龄(月)、身高(英寸)和体重(磅)的关系,研究者调查了 13 名学生的有关资料,数据列在表 11-5 中。问:将年龄化为相等后,就身高和体重两项指标整体而言,男生和女生差别有无统计学意义?

表 11-5 13 名学生的年龄(月)、身高(英寸)和体重(磅)

| 编号 ID | 性别 (SEX) | 年龄 (AGE) | 身高 (HEIGHT) | 体重 (WEIGHT) |
|-------|----------|----------|-------------|-------------|
| 1 | 1 | 164 | 66.5 | 112.0 |
| 2 | 1 | 189 | 65.0 | 114.0 |
| ... | ... | ... | ... | ... |
| 12 | 2 | 185 | 59.0 | 104.0 |
| 13 | 2 | 142 | 56.5 | 69.0 |

【例 11-6】 采用亚甲蓝治疗腰痛,用生理盐水作为对照,选取 71 例符合入选条件的患者进行试验,随机分为两组,其中试验组 36 例,对照组 35 例,分别于治疗前及疗后 3 个月时对每一位患者进行 VAS 评分和 ODI 评分,以此二项评分作为疗效评价的主要指标。数据如下表所示。请选择合适的统计学方法对资料进行分析,以评价亚甲蓝治疗腰痛的疗效。

表 11-6 两组腰痛患者的 VAS 评分和 ODI 评分

| 药物 | 编号 | ODI 评分 | | | |
|------|-----|--------|---------|-----|---------|
| | | 疗前 | 疗后 3 个月 | 疗前 | 疗后 3 个月 |
| 生理盐水 | 1 | 7.1 | 6.9 | 48 | 47 |
| | 2 | 6.3 | 6.4 | 39 | 41 |
| | ... | ... | ... | ... | ... |
| | 34 | 5.8 | 6.1 | 42 | 39 |
| | 35 | 6.3 | 5.5 | 47 | 42 |
| 亚甲蓝 | 36 | 8.0 | 2.0 | 45 | 7 |
| | 37 | 8.2 | 0.0 | 49 | 6 |
| | ... | ... | ... | ... | ... |
| | 70 | 6.2 | 2.1 | 46 | 12 |
| | 71 | 5.8 | 1.9 | 44 | 11 |

11.1.2 对数据结构的分析

例 11-1 资料为 5 名高血脂患者经益寿宁治疗后的胆固醇下降值与三酰甘油下降值。若 5 名高血脂患者在病情等方面的情况差不多,即具有良好的同质性,则本例资料属于单组设计二元定量资料。

例 11-2 资料为 20 名正常人分别于早间和下午两个时段测得的低频(LF)及高频(HF)心电频谱值,资料仅涉及一个实验因素,即“时段”,其有两个水平,即“早间”和“下午”。由于每名受试对象都要分别于两个时段重复测量两项相同的指标,故本例资料所对应的实验设计类型为配对设计。本例资料属于配对设计二元定量资料。

例 11-3 资料为 30 名初生到 3 周岁男、女婴幼儿的身高、体重和体表面积测量值,资料仅涉及一个实验因素,即“性别”,其有“男”、“女”2 个水平,而且,此实验因素为分组因素,因此,本例资料所对应的实验设计类型为单因素两水平设计(即成组设计)。本例资料属于单因素两水平设计三元定量资料。

例 11-4 资料所对应的实验设计类型与例 11-3 资料完全类似,只不过本例资料所涉及的实验分组因素“成绩水平”有 3 个水平而已。因此,本例资料属于单因素 3 水平设计二元定量资料,两个定量的指标分别为提转敏捷度评分与目标追踪评分。

例 11-5 资料为 13 名不同性别学生的年龄、身高和体重,属于单因素两水平设计三元定量资料。

例 11-6 资料涉及两个实验因素,即“药物种类”和“时间”,其中前者为实验分组因素,后者为重复测量因素,因此,从实验设计类型来看,本例资料属于具有一个重复测量的两因素设计。而且,本例资料有两项定量的指标,即“VAS 评分”和“ODI 评分”,因此,本例资料属于具有一个重复测量的两因素设计二元定量资料。

11.1.3 分析目的与统计分析方法选择

例 11-1 资料的分析目的是:考察“胆固醇下降值”与“三酰甘油下降值”两个指标的“总体均值向量(μ_1, μ_2)”是否等于“标准均值向量(0, 0)”。根据本例资料所对应的实验设计类型及数据结构,并假定资料满足参数检验的前提条件,可对本例资料进行单组设计定量资料的二元方差分析。

例 11-2 资料的分析目的是:考察正常人在“早间”和“下午”两时段的频谱值差别有无统计学意义。因此,根据本例资料所对应的实验设计类型及数据结构,并假定资料满足参数检验的前提条件,可采用配对设计定量资料的二元方差分析来处理。

例 11-3 资料的分析目的是:考察初生到 3 周岁婴幼儿中,男孩和女孩在三项指标上总的来说差异有无统计学意义。因此,根据本例资料所对应的实验设计类型及数据结构,并假定资料满足参数检验的前提条件,可对本例资料采用单因素两水平设计定量资料的三元方差分析来处理。

例 11-4 资料的分析目的是:比较三个不同水平上的小学生的两项定量指标的总体均值是否差不多,即差别有无统计学意义。因此,根据本例资料所对应的实验设计类型及数据结构,并假定资料满足参数检验的前提条件,可采用单因素 3 水平设计定量资料的二元方差分析来处理。

例 11-5 资料的分析目的是:考察将年龄化为相等后,就身高和体重两项指标整体而言,男生和女生的差别有无统计学意义。因此,根据本例资料所对应的实验设计类型及数据结构,并假定资料满足参数检验的前提条件,可采用单因素两水平设计二元定量资料的一元协方差分析来处理。这里的“二元”是指“身高”和“体重”两个统计指标,即结果变量;而“一元”是指协变量“年龄”。

例 11-6 资料的分析目的是:评价亚甲蓝治疗腰痛的疗效。要正确评价亚甲蓝治疗腰痛的疗效,就必须将试验组与对照组相比较。而两个组相比较时就必须考虑两个组的可比性。要使两个组具有良好的可比性,除了要求试验设计遵循均衡原则之外,还要求在统计分析阶段充分利用统计分析技术尽可能地消除那些定量的干扰因素对结果变量的影响。对于本例,由专业知识可知,治疗前患者的 VAS 评分和 ODI 评分不同,治疗相同时间后的疗效也会不同。因此,要正确地评价亚甲蓝治疗腰痛的疗效,在进行统计分析时宜将治疗前的 VAS 评分和 ODI 评分作为协变量,对资料进行单因素两水平设计二元定量资料的二元协方差分析,以消除因两个组治疗前 VAS 评分和 ODI 评分不同而对结果变量造成的影响。

11.2 单因素设计定量资料多元方差和协方差分析

11.2.1 对例 11-1 资料进行单组设计定量资料二元方差分析

所需的 SAS 程序如下, 程序名为 SASTJFX11_1.SAS。程序中 Y1 和 Y2 是变换后的变量, 等号右边的两个“0”代表两个指标的标准值, 在处理实际问题时, 可根据具体情况换成特定问题中的标准值。

```
DATA SASTJFX11_1;
INPUT X1 X2 @@ ;
Y1 = X1 - 0;
Y2 = X2 - 0;
CARDS;
  16   -4
  21   46
  57  -40
 -20  107
  17   86
;
RUN;
```

```
ODS HTML FILE = "D:\OUTPUT11_1.HTML";
PROC MEANS;
VAR Y1 - Y2;
ODS HTML;
RUN;
PROC GLM;
MODEL Y1 - Y2 = / SS3 NOUNI;
MANOVA H = INTERCEPT;
RUN;
QUIT;
ODS HTML CLOSE;
```

【主要分析结果及解释】

The MEANS Procedure

| Variable | N | Mean | Std Dev | Minimum | Maximum |
|----------|---|------------|------------|-------------|-------------|
| Y1 | 5 | 18.2000000 | 27.2891920 | -20.0000000 | 57.0000000 |
| Y2 | 5 | 39.0000000 | 61.1800621 | -40.0000000 | 107.0000000 |

Multivariate Analysis of Variance

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Intercept Effect

H = Type III SSCP Matrix for Intercept

E = Error SSCP Matrix

S = 1 M = 0 N = 0.5

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.13127792 | 9.93 | 2 | 3 | 0.0476 |

结果解释与专业结论: Y1 和 Y2 的样本均值分别为 18.2 和 39.0, 它们都大于 0, 而且, 其总体均值向量与标准均值向量(0, 0)之间的差别有统计学意义, 因为 Wilks' $\lambda = 0.1313$, 由其转换而来的近似 $F = 9.93$ 、 $P = 0.0476$ 。说明就所给的两个指标整体而言, 益寿宁有降血脂的作用。

11.2.2 对例 11-2 资料进行配对设计定量资料二元方差分析

所需要的 SAS 程序如下, 程序名为 SASTJFX11_2.SAS。需事先将数据以文本文件的形式保存于 D 盘 SASTJFX 文件夹内, 文件名为 data11_2.txt。

```
DATA SASTJFX11_2;
INFILE "D:\SASTJFX\data11_2.txt";
INPUT X1 X2 Y1 Y2;
D1 = X1 - X2; D2 = Y1 - Y2;
ODS HTML FILE = "D:\OUTPUT11_2.HTML";
PROC MEANS;
VAR X1 X2 Y1 Y2 D1 D2;
ODS HTML;
RUN;

PROC GLM;
MODEL D1 - D2 = /SS3 NOUNI;
MANOVA H = INTERCEPT;
RUN;
QUIT;
ODS HTML CLOSE;
```

【主要分析结果及解释】

| The MEANS Procedure | | | | | |
|---------------------|----|-------------|-------------|--------------|-------------|
| Variable | N | Mean | Std Dev | Minimum | Maximum |
| X1 | 20 | 154.4150000 | 94.6726342 | 37.6000000 | 367.9000000 |
| X2 | 20 | 144.4400000 | 68.0350869 | 22.8000000 | 326.3000000 |
| Y1 | 20 | 49.6600000 | 39.8821210 | 5.8000000 | 160.9000000 |
| Y2 | 20 | 105.7300000 | 80.6047870 | 22.9000000 | 343.4000000 |
| D1 | 20 | 9.9750000 | 111.9844250 | -183.6000000 | 307.9000000 |
| D2 | 20 | -56.0700000 | 79.9372655 | -313.6000000 | 0.8000000 |

Multivariate Analysis of Variance

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Intercept

Effect H = Type III SSCP Matrix for Intercept

E = Error SSCP Matrix

S = 1 M = 0 N = 8

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.59481144 | 6.13 | 2 | 18 | 0.0093 |

结果解释与专业结论：D1 和 D2 的样本均值分别为 9.975 和 -56.07，它们都不等于 0，而且，其总体均值向量与标准均值向量 (0, 0) 之间的差别有统计学意义，因为 Wilks' $\lambda = 0.5948$ ，由其转换而来的近似 $F = 6.13$ 、 $P = 0.0093$ 。这表明：对低、高频心电频谱两个指标整体而言，早间与下午两时段的频谱值是不相同的。对于低频心电频谱，从现有资料来看，早间和下午的频谱均值分别为 154.4 和 144.4，早间高于下午；对于高频心电频谱，从现有资料来看，早间和下午的频谱均值分别为 49.7 和 105.7，早间低于下午。

11.2.3 对例 11-3 资料进行单因素 2 水平设计定量资料三元方差分析

用“SEX”表示性别，男性用“1”表示、女性用“0”表示。三个观测指标分别用 X1、X2 和 Y 表示。以每个受试者的性别代码(0 或 1)和表 11-3 中后三列数据建立一个名为 DATA11_3.TXT 的文本文件，存放在 D 盘 SASTJFX 文件夹内，所需要的 SAS 程序如下，程序名为 SASTJFX11_3.SAS。

```
DATA SASTJFX11_3;
INFILE 'D:\SASTJFX\DATA11_3.TXT';
INPUT SEX X1 - X2 Y @@ ;
RUN;
ODS HTML FILE = "D:\OUTPUT11_3.HTML";

PROC GLM;
CLASS SEX;
MODEL X1 - X2 Y = SEX/SS3 NOUNI;
MANOVA H = SEX;
MEANS SEX;
RUN;QUIT;
ODS HTML CLOSE;
```

【主要分析结果及解释】

Multivariate Analysis of Variance

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall SEX Effect

H = Type III SSCP Matrix for SEX

E = Error SSCP Matrix

S = 1 M = 0.5 N = 12

| | Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|-----------|------------|---------|--------|--------|--------|
| Wilks' Lambda | | 0.87966309 | 1.19 | 3 | 26 | 0.3345 |

The GLM Procedure

| Level of | | X1 | | X2 | | Y | |
|----------|----|------------|------------|------------|------------|------------|------------|
| SEX | N | Mean | Std Dev | Mean | Std Dev | Mean | Std Dev |
| 0 | 15 | 73.1666667 | 16.9322881 | 8.12000000 | 4.40392682 | 3790.76000 | 1543.52371 |
| 1 | 15 | 75.2000000 | 18.3067123 | 8.58666667 | 4.80012897 | 4099.32667 | 1592.83788 |

结果解释与专业结论：三元方差分析的结果为 Wilks' $\lambda = 0.8797$ ，由其转换而来的近似 $F = 1.19$ 、 $P = 0.3345$ 。说明在初生到 3 周岁婴幼儿中，男孩和女孩在身高、体重和体表面积三项指标上总的来说并无明显差异。

11.2.4 对例 11-4 资料进行单因素 3 水平设计定量资料二元方差分析

用变量“SCORE”来表示成绩水平，成绩水平一般用“1”表示、较好用“2”表示、极好用“3”表示；定量的观测结果为 Y1 和 Y2。所需要的 SAS 程序如下，程序名为 SASTJFX11_4.SAS。数据文件名为 DATA11_4.TXT，存放在 D 盘 SASTJFX 文件夹内。

```
DATA SASTJFX11_4;
INFILE 'D:\SASTJFX\DATA11_4.TXT';
INPUT SCORE Y1 - Y2 @@ ;
RUN;
ODS HTML FILE = "D:\OUTPUT11_4.HT-
ML";
PROC GLM;
CLASS SCORE;
MODEL Y1 - Y2 = SCORE/NOUNI SS3;

CONTRAST 'SCORE1 vs SCORE2' SCORE1
-1 0;
CONTRAST 'SCORE1 vs SCORE3' SCORE1
0 -1;
CONTRAST 'SCORE2 vs SCORE3' SCORE0
1 -1;
MANOVA H = SCORE;
MEANS SCORE;
RUN;QUIT;
ODS HTML CLOSE;
```

【主要分析结果及解释】

Multivariate Analysis of Variance

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall SCORE Effect

H = Type III SSCP Matrix for SCORE

E = Error SSCP Matrix

S = 2 M = -0.5 N = 18.5

| | Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|-----------|------------|---------|--------|--------|--------|
| Wilks' Lambda | | 0.63971929 | 4.88 | 4 | 78 | 0.0015 |

MANOVA Test Criteria and Exact F Statistics for the
Hypothesis of No Overall SCORE1 vs SCORE2 Effect

H = Contrast SSCP Matrix for SCORE1 vs SCORE2

E = Error SSCP Matrix

S = 1 M = 0 N = 18.5

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.96732830 | 0.66 | 2 | 39 | 0.5232 |

MANOVA Test Criteria and Exact F Statistics for the
Hypothesis of No Overall SCORE1 vs SCORE3 Effect

H = Contrast SSCP Matrix for SCORE1 vs SCORE3

E = Error SSCP Matrix

S = 1 M = 0 N = 18.5

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.65101568 | 10.45 | 2 | 39 | 0.0002 |

MANOVA Test Criteria and Exact F Statistics for the
Hypothesis of No Overall SCORE2 vs SCORE3 Effect

H = Contrast SSCP Matrix for SCORE2 vs SCORE3

E = Error SSCP Matrix

S = 1 M = 0 N = 18.5

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.75712021 | 6.26 | 2 | 39 | 0.0044 |

The GLM Procedure

| Level of SCORE | N | y1 | | y2 | |
|-------------------|----|------------|------------|------------|------------|
| | | Mean | Std Dev | Mean | Std Dev |
| 1 | 16 | 57.0000000 | 9.35236156 | 92.625000 | 18.4133104 |
| 2 | 16 | 60.0000000 | 8.20568908 | 97.125000 | 19.1898411 |
| 3 | 11 | 70.0000000 | 6.40312424 | 113.454545 | 21.9014321 |

结果解释与专业结论: 总的来讲, 成绩水平不同的 9~11 岁年龄段小学生的两项神经行为功能指标得分是有差别的, 因为多元统计分析的结果为: Wilks' $\lambda = 0.6397$, 由其转换而来的近似 $F = 4.88$ 、 $P = 0.0015$ 。两两比较的结果为: 成绩水平一般的 9~11 岁年龄段小学生与成绩水平较好的 9~11 岁年龄段小学生, 其两项神经行为功能指标得分无明显差别 (Wilks' $\lambda = 0.9673$, $F = 0.66$, $P = 0.5232$), 成绩水平一般的 9~11 岁年龄段小学生与成绩水平极好的 9~11 岁年龄段小学生, 其两项神经行为功能指标得分有差别 (Wilks' $\lambda = 0.6510$, $F = 10.45$, $P = 0.0002$), 成绩水平较好的 9~11 岁年龄段小学生与成绩水平极好的 9~11 岁年龄段小学生, 其两项神经行为功能指标得分也有差别 (Wilks' $\lambda = 0.7571$, $F = 6.26$, $P = 0.0044$)。分析结果表明: 成绩水平一般的 9~11 岁年龄段小学生, 其提转敏捷度、目标追踪两项神经行为功能指标得分最低, 分别为 57.0 和 92.6; 成绩水平极好的 9~11 岁年龄段小学生, 其提转敏捷度、目标追踪两项神经行为功能指标得分最高, 分别为 70.0 和 113.5。

11.2.5 对例 11-5 资料进行单因素 2 水平设计二元定量资料的一元协方差分析

所需要的 SAS 程序如下, 程序名为 SASTJFX11_5.SAS。数据文件名为 DATA11_5.TXT, 存放在 D 盘 SASTJFX 文件夹内。

```
DATA SASTJFX11_5;
  INFILE 'D:\SASTJFX\DATA11_5.TXT';
  INPUT ID SEX AGE HEIGHT WEIGHT @@;
  RUN;
  ODS HTML FILE = "D:\OUTPUT11_5.HT-
  ML";
  PROC GLM;
  CLASS SEX;
  MODEL HEIGHT WEIGHT = AGE SEX AGE *
  SEX / SS3 NOUNI;
  MANOVA H = _ALL_;
  RUN;

PROC GLM;
  CLASS SEX;
  MODEL HEIGHT WEIGHT = AGE SEX
  /SS3 NOUNI;
  MANOVA H = _ALL_;
  LSMEANS SEX/CL STDERR PDIFF;
  RUN;
  QUIT;
  ODS HTML CLOSE;
```

【程序说明】 此程序共有两个 GLM 过程步。其中第一个过程步通过考察 AGE 与 SEX 之间是否存在交互作用来判断 SEX 取不同水平时, 所得到的响应变量与自变量之间的回归方程的斜率是否相等。若交互作用无统计学意义, 说明 SEX 取不同水平时, 所得到的回归方程斜率相同。MANOVA 语句中的参数 _ALL_ 代表了该语句前面 MODEL 语句右边的全部效应项。第二个 GLM 过程步是真正用于协方差分析的过程步, 其运行结果只有在第一个 GLM 过程步得出各协变量与定性的自变量之间的交互作用无统计学意义的前提下才有效。该过程步中, MODEL 语句右边列出的效应项不包含交互作用效应项。LSMEANS 语句用于在消除定量影响因素的干扰后, 比较定性自变量的不同水平所对应的总体响应的差别有无统计学意义。

【主要分析结果及解释】

以下是第一个过程步输出的结果。此过程步主要用来考察协方差分析应用的前提条件之一: 各组由自变量推测响应变量的回归方程的总体回归系数相等。所以, 其他对于主效应的分析可不必在意, 仅注意协变量与其他自变量的交互效应有无统计学意义即可。

Multivariate Analysis of Variance

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall AGE Effect

H = Type III SSCP Matrix for AGE

E = Error SSCP Matrix

S = 1 M = 0 N = 3

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.69445352 | 1.76 | 2 | 8 | 0.2326 |

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall SEX Effect

H = Type III SSCP Matrix for SEX

E = Error SSCP Matrix

S = 1 M = 0 N = 3

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.95299407 | 0.20 | 2 | 8 | 0.8248 |

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall AGE * SEX Effect

H = Type III SSCP Matrix for AGE * SEX

E = Error SSCP Matrix

S = 1 M = 0 N = 3

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.94683829 | 0.22 | 2 | 8 | 0.8037 |

分析结果表明,交互效应 AGE * SEX 对 HEIGHT 和 WEIGHT 两个响应变量整体的影响无统计学意义(Wilks' $\lambda = 0.947$, $F = 0.22$, $P = 0.8037 > 0.05$),所以本资料可进行协方差分析。

以下是第二个过程步输出的结果(省略掉了部分与第一个过程步输出结果相同的内容),该结果为最终的协方差分析结果。

Multivariate Analysis of Variance

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall AGE Effect

H = Type III SSCP Matrix for AGE

E = Error SSCP Matrix

S = 1 M = 0 N = 3.5

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.64714768 | 2.45 | 2 | 9 | 0.1411 |

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall SEX Effect

H = Type III SSCP Matrix for SEX

E = Error SSCP Matrix

S = 1 M = 0 N = 3.5

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.77801529 | 1.28 | 2 | 9 | 0.3232 |

The GLM Procedure

Least Squares Means

| SEX | HEIGHT LSMEAN | Standard Error | H0:LSMEAN = 0 | H0:LSMean1 = LSMean2 |
|-----|---------------|----------------|---------------|----------------------|
| | | | Pr > t | Pr > t |
| 1 | 62.6307144 | 1.1643251 | < .0001 | 0.1516 |
| 2 | 60.3183035 | 0.9184818 | < .0001 | |

| SEX | HEIGHT LSMEAN | 95% Confidence Limits | |
|-----|---------------|-----------------------|-----------|
| 1 | 62.630714 | 60.036436 | 65.224992 |
| 2 | 60.318303 | 58.271798 | 62.364809 |

Least Squares Means for Effect SEX

| i | j | Difference Between Means | 95% Confidence Limits for LSMean(i) - LSMean(j) | |
|---|---|--------------------------|---|----------|
| 1 | 2 | 2.312411 | -1.006587 | 5.631409 |

| SEX | HEIGHT LSMEAN | Standard Error | H0:LSMEAN = 0 | H0:LSMean1 = LSMean2 |
|-----|---------------|----------------|---------------|----------------------|
| | | | Pr > t | Pr > t |
| 1 | 110.491002 | 6.519381 | < .0001 | 0.2124 |
| 2 | 99.380624 | 5.142836 | < .0001 | |

| SEX | HEIGHT LSMEAN | 95% Confidence Limits | |
|-----|---------------|-----------------------|------------|
| 1 | 110.491002 | 95.964916 | 125.017087 |
| 2 | 99.380624 | 87.921672 | 110.839576 |

| Least Squares Means for Effect SEX | | | | |
|------------------------------------|---|--------------------------|---|-----------|
| Least Squares Means for Effect SEX | | | | |
| i | j | Difference Between Means | 95% Confidence Limits for LSMean(i) - LSMean(j) | |
| 1 | 2 | 11.110378 | -7.473616 | 29.694372 |

结果解释：经单因素 2 水平设计定量资料的一元协方差分析发现：就身高和体重两项指标整体而言，男生和女生并无差别（Wilks' $\lambda = 0.7780$, $F = 1.28$, $P = 0.3232 > 0.05$ ）。此外，年龄对身高和体重两项指标整体的影响无统计学意义（Wilks' $\lambda = 0.6471$, $F = 2.45$, $P = 0.1411 > 0.05$ ）。

专业结论：将年龄化为相等后，就身高和体重两项指标整体而言，男生和女生之间没有差别。

11.2.6 对例 11-6 资料进行单因素 2 水平设计二元定量资料的二元协方差分析

所需要的 SAS 程序如下，程序名为 SASTJFX11_6.SAS。数据文件名为 DATA11_6.TXT，存放在 D 盘 SASTJFX 文件夹内。

```
DATA SASTJFX11_6;
INFILE 'D:\SASTJFX\DATA11_6.TXT';
INPUT MEDICINE ID VAS_BL VAS
      ODI_BL ODI @@ ;
RUN;
ODS HTML FILE = "D:\OUTPUT11_6.HTML";
PROC GLM;
CLASS MEDICINE;
MODEL VAS ODI = VAS_BL ODI_BL MEDICINE / SS3 NOUNI;
MANOVA H = _ALL_;
RUN;

PROC GLM;
CLASS MEDICINE;
MODEL VAS ODI = VAS_BL ODI_BL MEDICINE / SS3 NOUNI;
MANOVA H = _ALL_;
RUN;
```

【程序说明】 程序中用“MEDICINE”表示药物种类（0 = 亚甲蓝，1 = 生理盐水），用“VAS_BL”和“ODI_BL”分别表示治疗前的 VAS 和 ODI 评分，即 VAS 和 ODI 评分的基线值。其他说明同例 11-5。

【主要分析结果及解释】

以下是第一个过程步输出的初步的二元定量资料的二元协方差分析结果。此过程步主要用来考察协方差分析应用的前提条件之一：各组由自变量推测响应变量的回归方程的总体回归系数相等。所以，其他对于主效应的分析可不必在意，仅注意协变量与其他自变量的交互效应有无统计学意义即可。

Multivariate Analysis of Variance

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall VAS_BL Effect

H = Type III SSCP Matrix for VAS_BL

E = Error SSCP Matrix

S = 1 M = 0 N = 31

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.94513739 | 1.86 | 2 | 64 | 0.1644 |

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall ODI_BL Effect

H = Type III SSCP Matrix for ODI_BL

E = Error SSCP Matrix

S = 1 M = 0 N = 31

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.92309714 | 2.67 | 2 | 64 | 0.0773 |

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall MEDICINE Effect

H = Type III SSCP Matrix for MEDICINE

E = Error SSCP Matrix

S = 1 M = 0 N = 31

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.97112342 | 0.95 | 2 | 64 | 0.3915 |

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall VAS_BL * MEDICINE Effect

H = Type III SSCP Matrix for VAS_BL * MEDICINE

E = Error SSCP Matrix

S = 1 M = 0 N = 31

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.97872984 | 0.70 | 2 | 64 | 0.5026 |

分析结果表明,交互效应 VAS_BL * MEDICINE 对 VAS 和 ODI 两个响应变量整体的影响无统计学意义(Wilks' $\lambda = 0.979$, $F = 0.70$, $P = 0.5026 > 0.05$)。

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall ODI_BL * MEDICINE Effect

H = Type III SSCP Matrix for ODI_BL * MEDICINE

E = Error SSCP Matrix

S = 1 M = 0 N = 31

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.93233222 | 2.32 | 2 | 64 | 0.1062 |

分析结果表明,交互效应 ODI_BL * MEDICINE 对 VAS 和 ODI 两个响应变量整体的影响也无统计学意义(Wilks' $\lambda = 0.932$, $F = 2.32$, $P = 0.1062 > 0.05$),所以本资料可进行协方差分析。

以下是第二个过程步输出的二元定量资料的二元协方差分析结果,该结果为最终结果。

Multivariate Analysis of Variance

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall VAS_BL Effect

H = Type III SSCP Matrix for VAS_BL

E = Error SSCP Matrix

S = 1 M = 0 N = 32

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.95056817 | 1.72 | 2 | 66 | 0.1877 |

分析结果表明：VAS_BL 对 VAS 和 ODI 两个响应变量整体的影响无统计学意义 (Wilks' $\lambda = 0.951$, $F = 1.72$, $P = 0.1877 > 0.05$)。

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall ODI_BL Effect

H = Type III SSCP Matrix for ODI_BL

E = Error SSCP Matrix

S = 1 M = 0 N = 32

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.89094715 | 4.04 | 2 | 66 | 0.0221 |

分析结果表明：ODI_BL 对 VAS 和 ODI 两个响应变量整体的影响有统计学意义 (Wilks' $\lambda = 0.891$, $F = 4.04$, $P = 0.0221 < 0.05$)。

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall MEDICINE Effect

H = Type III SSCP Matrix for MEDICINE

E = Error SSCP Matrix

S = 1 M = 0 N = 32

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.27950690 | 85.07 | 2 | 66 | <.0001 |

分析结果表明：MEDICINE 对 VAS 和 ODI 两个响应变量整体的影响有统计学意义 (Wilks' $\lambda = 0.280$, $F = 85.07$, $P < 0.0001$)。

The GLM Procedure

| Level of MEDICINE | N | VAS | | ODI | | VAS_BL | | ODI_BL | |
|----------------------|----|------------|------------|------------|------------|------------|------------|------------|------------|
| | | Mean | Std Dev | Mean | Std Dev | Mean | Std Dev | Mean | Std Dev |
| 0 | 36 | 2.49444444 | 1.73829875 | 16.0000000 | 11.9139774 | 7.23333333 | 1.23519808 | 48.4722222 | 5.12409496 |
| 1 | 35 | 6.35142857 | 1.16628405 | 48.4000000 | 7.7694727 | 6.73142857 | 1.16158397 | 49.3714286 | 6.79136872 |

这是未根据协变量校正前两药物组响应变量 VAS 和 ODI 的均值和标准差，以及协变量 VAS_BL 和 ODI_BL 的均值和标准差。

The GLM Procedure

Least Squares Means

| MEDICINE | VAS LSMEAN | Standard Error | H0: LSMEAN = 0 | H0: LSMEAN1 = LSMEAN2 |
|----------|------------|----------------|----------------|-----------------------|
| | | | Pr > t | Pr > t |
| 0 | 2.49181949 | 0.24159435 | <.0001 | <.0001 |
| 1 | 6.35412852 | 0.24526504 | <.0001 | |

这是根据协变量校正后两药物组响应变量 VAS 的均值和标准差，以及对两药物组响应变量 VAS 的总体均值进行比较的结果：两药物组校正后的 VAS 评分差异有统计学意义 ($P < 0.0001$)。

| MEDICINE | VAS LSMEAN | 95% Confidence Limits | |
|----------|------------|-----------------------|----------|
| 0 | 2.491819 | 2.009595 | 2.974044 |
| 1 | 6.354129 | 5.864577 | 6.843680 |

这是两药物组响应变量 VAS 的校正均值及其 95% 置信限。可以看出：两个组的 VAS 的 95% 置信区间不重叠，表明两药物组校正后的 VAS 评分差异有统计学意义，而且，试验药组 (MEDICINE = 0) VAS 的校正均值为 2.491819，小于对照组 VAS 的校正均值 6.354129。

Least Squares Means for Effect MEDICINE

Least Squares Means for Effect SEX

| i | j | Difference Between Means | 95% Confidence Limits for LSMean(i) - LSMean(j) | |
|---|---|--------------------------|---|-----------|
| 1 | 2 | -3.862309 | -4.573305 | -3.151313 |

这是两药物组响应变量 VAS 的校正均值的差值及其 95% 置信限。试验药组与对照组 VAS 的校正均值之差值为 -3.86, 差值的 95% 置信区间约为 -4.6 ~ -3.2, 置信区间不包含 0。

| MEDICINE | VAS LSMEAN | Standard Error | H0: LSMEAN = 0 Pr > t | H0: LSMEAN1 = LSMEAN2 Pr > t |
|----------|------------|----------------|---------------------------|----------------------------------|
| 0 | 16.5807892 | 1.6331914 | <.0001 | <.0001 |
| 1 | 47.8026168 | 1.6580055 | <.0001 | |

这是根据协变量校正后两药物组响应变量 ODI 的均值和标准差, 以及对两药物组响应变量 ODI 的总体均值进行比较的结果: 两药物组校正后的 ODI 评分差异有统计学意义 (P < 0.0001)。

| MEDICINE | VAS LSMEAN | 95% Confidence Limits | |
|----------|------------|-----------------------|-----------|
| 0 | 16.580789 | 13.320926 | 19.840653 |
| 1 | 47.802617 | 44.493224 | 51.112010 |

这是两药物组响应变量 ODI 的校正均值及其 95% 置信限。可以看出: 两个组的 ODI 的 95% 置信区间不重叠, 表明两药物组校正后的 ODI 评分差异有统计学意义, 而且, 试验药组 (MEDICINE = 0) ODI 的校正均值为 16.580789, 小于对照组 ODI 的校正均值 47.802617。

Least Squares Means for Effect MEDICINE

Least Squares Means for Effect SEX

| i | j | Difference Between Means | 95% Confidence Limits for LSMean(i) - LSMean(j) | |
|---|---|--------------------------|---|------------|
| 1 | 2 | -31.221828 | -36.028202 | -26.415453 |

这是两药物组响应变量 ODI 的校正均值的差值及其 95% 置信限。试验药组与对照组 ODI 的校正均值之差值为 -31.22, 差值的 95% 置信区间约为 -36.0 ~ -26.4, 置信区间不包含 0。

专业结论: 亚甲蓝有治疗腰痛的短期疗效。治疗 6 个月时间, VAS 评分可降低 3.2 ~ 4.6 分, ODI 评分可降低 26.4 ~ 36.0 分。

11.3 本章小结

本章结合实例介绍了用 SAS 实现单组设计、配对设计、单因素 2 水平设计、单因素多水平设计四种单因素设计定量资料的多元方差分析方法及单因素 2 水平、单因素多水平设计多元定量资料的一元或多元协方差分析方法。本章的所有例题均给出了 SAS 程序、分析结果、统计学结论和专业结论, 读者在解决类似的实际问题时完全可以如法炮制。当然, 要使资料的分析正确, 读者还必须注意以下几点: (1) 检查试验设计或调查设计的正确性、合理性; (2) 检查资料是否为多元定量资料, 是否需要做多元统计分析; (3) 检查多元定量资料是否满足多元方差分析或一元、多元协方差分析的前提条件。有关多元方差分析或一元、多元协方差分析的前提条件的检查方法比较复杂, 本章未予介绍, 请读者查阅相关文献。

(周诗国)

第 12 章 多因素设计多元定量资料差异性分析

第 11 章已经介绍了几种单因素设计定量资料的多元方差分析或协方差分析的 SAS 实现方法，本章将接着介绍用 SAS 实现几种常用多因素设计定量资料的多元方差分析或协方差分析方法，所涉及的实验设计类型主要是随机区组设计(也称配伍组设计)、析因设计、含区组因素的析因设计、正交设计和重复测量设计。

12.1 问题、数据及统计分析方法的选择

12.1.1 问题与数据

【例 12-1】 某研究者进行了一项试验，研究长期饲喂高锌日粮对断奶仔猪免疫机能的影响，探讨高锌对断奶仔猪的毒副作用。从 10 窝 23 日龄 ± 1 日龄“杜长大”三元杂交断奶仔猪中选择临床检查健康的仔猪 60 头(每窝选 6 头，雌、雄各半)，根据窝别和性别将其分为 20 个区组，每个区组内 3 只条件接近或相似的仔猪。然后，将区组内的 3 只仔猪随机分配到 A、B、C 三个处理组中去。其中，A 组仔猪饲喂基础日粮，B 组和 C 组仔猪分别饲喂基础日粮 + 3000mg/kg 氧化锌和基础日粮 + 500mg/kg 氧化锌，试验期为 70d。试验仔猪采用群饲，自由采食，自由饮水，免疫消毒程序按猪场常规方法进行。断奶后第 70d 前腔静脉采血检测血清 IgG、IgA 和 IgM 的水平，数据如表 12-1 所示(部分数据省略)。问：3 个组仔猪的血清 IgG、IgA 和 IgM 水平差别有无统计学意义？

表 12-1 断奶后第 70d 时各组仔猪的血清 IgG、IgA 和 IgM 水平 (g/L)

| 区组 | IgG 水平 | | | IgA 水平 | | | IgM 水平 | | |
|-----|--------|-------|-------|--------|-------|-------|--------|-------|-------|
| | A | B | C | A | B | C | A | B | C |
| 1 | 0.383 | 0.222 | 0.211 | 0.138 | 0.078 | 0.090 | 0.325 | 0.273 | 0.218 |
| 2 | 0.465 | 0.224 | 0.179 | 0.137 | 0.069 | 0.092 | 0.276 | 0.214 | 0.218 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 19 | 0.415 | 0.185 | 0.198 | 0.147 | 0.087 | 0.086 | 0.269 | 0.292 | 0.170 |
| 20 | 0.471 | 0.193 | 0.268 | 0.121 | 0.097 | 0.083 | 0.310 | 0.235 | 0.184 |

【例 12-2】 欲研究制造塑料胶片过程中挤压度和添加剂剂量对其性能的影响，随机抽样获得 20 个塑料胶片样品，按 2×2 析因设计，挤压度和添加剂剂量各选 2 个水平，塑料胶片的性能用 3 个指标表达，即抗泪度、光泽度和混浊度，数据见表 12-2。问：塑料胶片制造过程中挤压度和添加剂剂量对其性能的影响有无统计学意义？

表 12-2 挤压度和添加剂剂量对塑料胶片的影响

| 挤压度 | 添加剂剂量 | 抗泪度 | 光泽度 | 混浊度 | 挤压度 | 添加剂剂量 | 抗泪度 | 光泽度 | 混浊度 |
|-----|-------|-----|-----|-----|-----|-------|-----|-----|-----|
| 1 | 1 | 6.5 | 9.5 | 4.4 | 2 | 1 | 6.7 | 9.1 | 2.8 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 1 | 1 | 6.5 | 9.2 | 0.8 | 2 | 1 | 6.8 | 8.5 | 3.4 |
| 1 | 2 | 6.9 | 9.1 | 5.7 | 2 | 2 | 7.1 | 9.2 | 8.4 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 1 | 2 | 6.3 | 9.4 | 5.7 | 2 | 2 | 7.6 | 9.2 | 1.9 |

【例 12-3】 有人研究浸泡消毒藻酸钾印模对模型尺寸稳定性的影响，假定实验是这样做的：考虑“消毒剂种类”和“消毒时间”两个实验因素，前者有“2% 强化中性戊二醛”、“0.5% 次氯酸”和“0.1% 碘伏”3 个水平，后者有“10 分钟”、“20 分钟”和“30 分钟”3 个水平。先制作 54 件藻酸钾印模试件，虽然每件试件的制作都在相同的温度下按标准方法调拌印模材料，而且所用量具相同，调拌速度相同，但为了加快制模速度，缩短制模时间，由 6 位模型制作专业技术人员各自制作完成 9 个完好的印模试件，每位技术员制作的印模试件都编上号。印模试件制作完成后，根据“消毒剂种类”和“消毒时间”两个实验因素的水平组合数(3×3=9)，同时考虑到试件制作员的不同可能对实验带来的影响，对试件进行分组时分别将每位试件制作员所制作的 9 件试件随机分为 9 组。每一组试件接受一种消毒剂消毒一定时间后取出灌注模型。24h 后用测量显微镜测量各组模型 l6 近远中径(MD)，l6 颊舌径(LB)，6 l6 间距(CA)，l36 间距(AP)，线性长度值(mm)。测量结果如表 12-3、表 12-4、表 12-5 和表 12-6 所示。假定“消毒剂种类”和“消毒时间”两个实验因素对实验结果的影响地位平等，且已知此四项定量指标在专业上存在相互联系或影响，并假定资料满足参数检验的前提条件，试对此资料进行统计分析。

表 12-3 各组 l6 近远中径(MD)的测量结果

| 制模员 编号 | MD 长度值(mm) | | | | | | | | |
|-----------|-------------|-------|-------|--------|-------|-------|--------|-------|-------|
| | A1(B1 | B2 | B3) | A2(B1 | B2 | B3) | A3(B1 | B2 | B3) |
| 1 | 9.408 | 9.411 | 9.421 | 9.412 | 9.415 | 9.418 | 9.415 | 9.417 | 9.427 |
| 2 | 9.415 | 9.411 | 9.429 | 9.416 | 9.419 | 9.408 | 9.412 | 9.418 | 9.419 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 5 | 9.408 | 9.410 | 9.422 | 9.420 | 9.419 | 9.427 | 9.418 | 9.416 | 9.421 |
| 6 | 9.412 | 9.428 | 9.424 | 9.420 | 9.415 | 9.406 | 9.416 | 9.417 | 9.421 |

注：A1、A2、A3 分别表示 2% 强化中性戊二醛、0.5% 次氯酸和 0.1% 碘伏 3 种消毒剂；B1、B2、B3 分别表示 10min、20min 和 30min 3 种消毒时间，下同。

表 12-4 各组模型 l6 颊舌径(LB)的测量结果

| 制模员 编号 | LB 长度值(mm) | | | | | | | | |
|-----------|-------------|-------|-------|--------|-------|-------|--------|-------|-------|
| | A1(B1 | B2 | B3) | A2(B1 | B2 | B3) | A3(B1 | B2 | B3) |
| 1 | 8.988 | 8.994 | 9.011 | 8.979 | 8.986 | 9.008 | 8.980 | 8.986 | 8.999 |
| 2 | 8.983 | 8.991 | 9.018 | 8.975 | 8.972 | 9.012 | 8.979 | 8.984 | 9.014 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 5 | 8.978 | 8.986 | 9.008 | 8.980 | 8.989 | 9.014 | 8.987 | 8.991 | 9.008 |
| 6 | 8.984 | 8.997 | 9.004 | 8.975 | 8.972 | 9.012 | 8.986 | 8.986 | 9.008 |

表 12-5 各组模型 6 l6 间距(CA)的测量结果

| 制模员 编号 | CA 长度值(mm) | | | | | | | | |
|-----------|-------------|--------|--------|--------|--------|--------|--------|--------|--------|
| | A1(B1 | B2 | B3) | A2(B1 | B2 | B3) | A3(B1 | B2 | B3) |
| 1 | 51.377 | 51.391 | 51.398 | 51.378 | 51.378 | 51.401 | 51.368 | 51.379 | 51.385 |
| 2 | 51.384 | 51.391 | 51.409 | 51.384 | 51.385 | 51.407 | 51.383 | 51.381 | 51.411 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 5 | 51.376 | 51.389 | 51.391 | 51.373 | 51.389 | 51.412 | 51.375 | 51.393 | 51.428 |
| 6 | 51.382 | 51.375 | 51.399 | 51.386 | 51.403 | 51.424 | 51.383 | 51.382 | 51.404 |

表 12-6 各组模型 I36 间距 (AP) 的测量结果

| 制模员 编号 | AP 长度值 (mm) | | | | | | | | |
|-----------|-------------|--------|--------|---------|--------|--------|---------|--------|--------|
| | A1 (B1) | B2 | B3) | A2 (B1) | B2 | B3) | A3 (B1) | B2 | B3) |
| 1 | 29.543 | 29.569 | 29.549 | 29.531 | 29.541 | 29.558 | 29.541 | 29.535 | 29.548 |
| 2 | 29.536 | 29.560 | 29.546 | 29.529 | 29.555 | 29.544 | 29.530 | 29.535 | 29.543 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 5 | 29.527 | 29.550 | 29.546 | 29.527 | 29.525 | 29.533 | 29.521 | 29.532 | 29.546 |
| 6 | 29.527 | 29.550 | 29.544 | 29.524 | 29.555 | 29.552 | 29.536 | 29.533 | 29.545 |

【例 12-4】 航空煤油(以下简称航煤)由常压塔一线出来,由于含有环烷酸,需要进行碱洗与水洗。目前采用的工艺条件,使航煤质量不符合高标准要求。研究者通过试验,寻求影响航煤质量的主要因素及比较好的工艺条件,以提高航煤的洁净度。试验中考察了如下三个响应变量(见表 12-7)。

微水量(y_1): 小于 90ppm 为合格。

微量钠(y_2): 小于 0.5ppm 为合格。

酸度(y_3): 小于 0.7KOH mg/100ml 为合格。

以上三个响应变量,要求其取值越小越好。

表 12-7 所考察的试验因素及其水平取值

| 水平 代号 | 碱洗温度 (A, °C) | 碱洗沉降 时间 (B, min) | 碱洗浓度 (C, %) | 碱洗搅拌 速度 (D) | 水洗量 (E) | 水洗温度 (F, °C) | 水洗沉降时间 (G, min) | 水洗搅拌 速度 (H) |
|----------|-----------------|---------------------|----------------|----------------|------------|-----------------|--------------------|----------------|
| 1 | 40 | 90 | 2 | 中速 | 30 | 30 | 40 | 低速 |
| 2 | 50 | 30 | 6 | 高速 | 50 | 40 | 20 | 中速 |
| 3 | 30 | 60 | 4 | 低速 | 10 | 50 | 60 | 高速 |

试验的安排: 选用 $L_{27}(3^{13})$ 安排试验。表头设计如表 12-8 所示。

表 12-8 航煤试验的表头设计

| 列号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|--------------|---|---|-------|----------------|---|---|---|---|---|----|-------|----|----|
| 试验因素及其交互作用代号 | A | F | A * F | A * F E * H | B | C | H | | E | | E * H | D | G |

值得注意的是,在这个表头设计中, $A * F$ 与 $E * H$ 两个一级交互作用各有一部分同时出现在第 4 列上,即交互作用效应部分出现了混杂。由于从第 5 列到第 13 列中任取两列来安排 E 和 H 两个因素,其交互作用必有一部分会出现在前 4 列上,也就是说,在不增大正交表的前提下,在 $L_{27}(3^{13})$ 正交表上考察两个独立的一级交互作用必然会导致部分效应混杂。

试验结果如表 12-9 所示。假定资料满足参数检验的前提条件,试对此资料进行统计分析。

表 12-9 试验结果

| 试验号 | 各因素(列号)的水平代号 | | | | | | | | y_1 | y_2 | y_3 |
|-----|--------------|------|------|------|------|------|-------|-------|-------|-------|-------|
| | 1(A) | 2(F) | 5(B) | 6(C) | 7(H) | 9(E) | 12(D) | 13(G) | | | |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 71 | 0.25 | 0.25 |
| 2 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 93 | 0.70 | 0.28 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 26 | 3 | 3 | 2 | 1 | 3 | 3 | 2 | 1 | 121 | 0.22 | 0.22 |
| 27 | 3 | 3 | 3 | 2 | 1 | 1 | 3 | 2 | 105 | 0.37 | 1.06 |

注: 资料来源: 中国农科院研究生院远程教育网络: <http://www.gscaas.net.cn>。

【例 12-5】 某医生观察了 20 名正常人(HA)及 20 名稳定型心绞痛(SAP)患者的 24h 动态心电图,并分析出早晨 3 个小时各小时的低频心电频谱值(LF)、高频心电频谱值(HF),结果如表 12-10 所示。假定资料满足参数检验的前提条件,试对此资料进行统计分析。

表 12-10 20 名正常人及 20 名稳定型心绞痛患者的低、高频心电频谱值的自然对数值

| 受试对象 患病情况 | 受试对 象编号 | 低频谱值的自然对数值 | | | 高频谱值的自然对数值 | | |
|--------------|------------|------------|--------|--------|------------|--------|--------|
| | | 第 1 小时 | 第 2 小时 | 第 3 小时 | 第 1 小时 | 第 2 小时 | 第 3 小时 |
| 未患心绞痛 | 1 | 4.66 | 4.29 | 4.77 | 2.89 | 3.03 | 3.57 |
| | 2 | 4.54 | 4.69 | 4.58 | 2.65 | 2.77 | 3.04 |
| | ... | ... | ... | ... | ... | ... | ... |
| | 19 | 4.88 | 5.19 | 4.79 | 3.73 | 4.27 | 4.16 |
| | 20 | 5.33 | 5.82 | 5.88 | 4.19 | 4.41 | 4.51 |
| 稳定型心绞痛 | 21 | 3.83 | 6.64 | 6.10 | 3.46 | 4.97 | 5.06 |
| | 22 | 5.55 | 6.61 | 6.65 | 3.98 | 5.11 | 5.70 |
| | ... | ... | ... | ... | ... | ... | ... |
| | 39 | 4.99 | 4.61 | 5.02 | 3.73 | 3.35 | 3.38 |
| | 40 | 4.72 | 5.35 | 5.70 | 2.67 | 3.00 | 3.66 |

资料来源:柳青主编.中国医学统计百科全书·多元统计分册.2004:119.

【例 12-6】 假设研究者调查了 15 名沥青工和 12 名焦炉工的年龄、工龄、吸烟情况,并测得血清 P21、P53 水平及外周血淋巴细胞 SCE、染色体畸变数和染色体畸变细胞数如表 12-11 所示,试选取合适的统计学分析方法,分析工种、年龄、工龄和吸烟量对工人各项生物标志物水平的影响。

表 12-11 15 名沥青工和 12 名焦炉工的年龄、工龄、吸烟情况及血清 P21、P53 水平、外周血淋巴细胞 SCE、染色体畸变数和染色体畸变细胞数

| ID | 工种 | 年龄 (岁) | 工龄 (年) | 吸烟 情况 | 血清 P21 水平 | P53 水平 | SCE | 染色体畸 变数 | 染色体畸变 细胞数 |
|-----|-----|-----------|-----------|----------|--------------|-----------|-------|------------|--------------|
| 1 | 1 | 46 | 25 | 2 | 2138 | 0.35 | 8.11 | 4 | 4 |
| 2 | 1 | 35 | 12 | 3 | 3510 | 1.43 | 6.84 | 3 | 3 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 26 | 2 | 34 | 14 | 3 | 4322 | 0.41 | 15.00 | 5 | 5 |
| 27 | 2 | 50 | 32 | 3 | 2862 | 0.69 | 8.80 | 2 | 2 |

注:工种:1=沥青工,2=焦炉工;吸烟情况:1=不吸烟,2=少量吸烟,3=大量吸烟。

12.1.2 对数据结构的分析

例 12-1 资料属于随机区组设计三元定量资料。本例资料所涉及的实验因素有两个,即“氧化锌剂量”和“窝别”,前者为实验分组因素,后者为区组因素。受试对象为临床检查健康的仔猪 60 头(每窝选 6 头,雌、雄各半),根据窝别和性别将其分为 20 个区组,每个区组内的 3 只条件接近或相似。统计指标有 3 项,即仔猪的血清 IgG、IgA 和 IgM 水平。

例 12-2 资料显然属于 2×2 析因设计三元定量资料,两个实验因素分别为“挤压度”和“添加剂剂量”,三个定量的统计指标分别为抗泪度、光泽度和混浊度,受试对象为通过随机抽样所获得的 20 个塑料胶片样品。

例 12-3 资料涉及两个地位平等的实验分组因素,即“消毒剂种类”和“消毒时间”,基本上具备了析因设计的特点,由于又多了一个区组因素,即“制模员编号”,故本研究所采用的实验

设计类型为含区组因素的析因设计。由于本例资料涉及四个定量的观测指标，而且此四项定量的观测指标在专业上存在相互联系或影响，故本例资料为含区组因素的析因设计四元定量资料。

例 12-4 资料显然属于正交设计三元定量资料。此资料所涉及的实验因素有 8 个，分别为碱洗温度、碱洗沉降时间、碱洗浓度、碱洗搅拌速度、水洗量、水洗温度、水洗沉降时间、水洗搅拌速度；统计指标有 3 个，即微水量、微量钠、酸度，而且此 3 个指标需要同时予以考虑；受试对象为 27 份在不同条件下经过碱洗与水洗后的航空煤油样品。

例 12-5 资料涉及“患病情况”和“时间”两个实验因素，其中前者为实验分组因素，后者为重复测量因素。而且，此资料涉及两项定量的观测指标，即“低频谱值的自然对数”和“高频谱值的自然对数”。因此，本例资料属于具有一个重复测量的两因素设计二元定量资料。

例 12-6 资料为 15 名沥青工和 12 名焦炉工的年龄、工龄、吸烟情况、血清 P21 水平、P53 水平及外周血淋巴细胞 SCE、染色体畸变数和染色体畸变细胞数。至于此资料属于何种实验设计类型，需要根据统计分析目的来确定。

12.1.3 分析目的与统计分析方法的选择

例 12-1 资料的分析目的是：考察 3 个饲料组仔猪的血清 IgG、IgA 和 IgM 水平差别有无统计学意义。根据本例资料所对应的实验设计类型及数据结构，并假定资料满足参数检验的前提条件，可对本例资料进行随机区组设计定量资料的三元方差分析。

例 12-2 资料的分析目的是：考察塑料胶片制造过程中“挤压力”和“添加剂剂量”对其性能的影响有无统计学意义。因此，根据本例资料所对应的实验设计类型及数据结构，并假定资料满足参数检验的前提条件，可对本例资料采用两因素析因设计定量资料的三元方差分析。

例 12-3 资料的分析目的是：考察“消毒剂种类”和“消毒时间”两个实验因素对实验结果的影响。因此，根据本例资料所对应的实验设计类型及数据结构，并假定资料满足参数检验的前提条件，可对本例资料采用含区组因素的析因设计定量资料的四元方差分析。

例 12-4 资料的分析目的是：考察碱洗与水洗工艺条件对航空煤油洁净度的影响。因此，根据本例资料所对应的实验设计类型及数据结构，并假定资料满足参数检验的前提条件，可采用正交设计定量资料的三元方差分析。

例 12-5 资料的分析目的是：考察“患病情况”和“时间”对心电图的影响有无统计学意义。因此，根据本例资料所对应的实验设计类型及数据结构，并假定资料满足参数检验的前提条件，可对本例资料采用具有一个重复测量的两因素设计定量资料的二元方差分析。

例 12-6 资料的分析目的是：考察工种、年龄、工龄和吸烟量对工人各项生物标志物水平的影响有无统计学意义。由于本例资料涉及的自变量有年龄、工龄、工种和吸烟情况共四个，其中前两个为定量变量，后两个为定性变量；涉及的统计指标有血清 P21、P53 水平及外周血淋巴细胞 SCE、染色体畸变数和染色体畸变细胞数共五个，均为定量变量，且五个统计指标在专业上是有一定联系的。因此，要分析各自变量对五个统计指标的影响情况，宜采用两因素析因设计五元定量资料的二元协方差分析，其中的分组变量为工种和吸烟情况，协变量为年龄和工龄。

12.2 多因素设计定量资料多元方差和协方差分析

12.2.1 对例 12-1 资料进行随机区组设计定量资料三元方差分析

【程序说明】 分别用 block 和 dose 表示“区组”和“氧化锌剂量”，3 项观测指标分别用 IgG、IgA 和 IgM 表示。所需的 SAS 程序如下，程序名为 SASTJFX12_1.SAS。数据文件名为 DATA12_1.TXT，存放在 D 盘根目录下。

```
DATA SASTJFX12_1;
INFILE 'D:\DATA12_1.TXT';
DO block=1 TO 20;
DO dose=0,3000,500;
INPUT IgG IgA IgM @@ ;
OUTPUT;
END; END;
RUN;
ODS HTML FILE = "D:\OUTPUT12_1.HT-
ML";

PROC GLM DATA = SASTJFX12_1;
CLASS dose block;
MODEL IgG IgA IgM = dose block /SS3
NOUNI;
MANOVA H=dose block;
MEANS dose;
RUN;
QUIT;
ODS HTML CLOSE;
```

【主要分析结果及解释】

The GLM Procedure

Multivariate Analysis of Variance

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall dose Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.02820722 | 59.45 | 6 | 72 | <.0001 |

同时考察 IgG 水平、IgA 水平和 IgM 水平三个响应变量时，不同氧化锌剂量组之间的差别有统计学意义 (Wilks' $\lambda = 0.02820722$, $F = 59.45$, $P < 0.0001$)。

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall block Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.26311349 | 1.07 | 57 | 108.17 | 0.3730 |

同时考察 IgG 水平、IgA 水平和 IgM 水平三个响应变量时，区组之间的差别无统计学意义 (Wilks' $\lambda = 0.26311349$, $F = 1.07$, $P = 0.3730 > 0.05$)。

The GLM Procedure

| Level of dose | N | IgG | | IgA | | IgM | |
|------------------|----|------------|------------|------------|------------|------------|------------|
| | | Mean | Std Dev | Mean | Std Dev | Mean | Std Dev |
| 0 | 20 | 0.41160000 | 0.03949071 | 0.12425000 | 0.01831845 | 0.28430000 | 0.02297619 |
| 500 | 20 | 0.20830000 | 0.02743240 | 0.08100000 | 0.00780013 | 0.22730000 | 0.03227367 |
| 3000 | 20 | 0.18160000 | 0.03268091 | 0.07960000 | 0.00841302 | 0.24950000 | 0.02774223 |

这是因素 dose 各水平下的样本含量、均数和标准差。

专业结论：用基础日粮添加不同剂量氧化锌的饲料喂养断奶仔猪 70d 后，各剂量组仔猪的血清 IgG 水平、IgA 水平和 IgM 水平三项指标整体上会有差别。往基础日粮添加氧化锌会降低断奶仔猪的血清 IgG 水平、IgA 水平和 IgM 水平。

12.2.2 对例 12-2 资料进行两因素析因设计定量资料三元方差分析

【程序说明】 用 A 和 B 分别表示“挤压力”和“添加剂剂量”两个自变量，用 Y1、Y2 和 Y3 分别表示抗泪度、光泽度和混浊度三个指标变量。所需的 SAS 程序如下，程序名为 SAS-TJFX12_2.SAS。数据文件名为 DATA12_2.TXT，存放在 D 盘根目录下。

```
DATA SASTJFX12_2;
INFILE "D:\DATA12_2.TXT";
INPUT A B Y1 -Y3@@ ;RUN;
ODS HTML FILE="D:\OUTPUT12_2.HTML";
PROC GLM;
CLASS A B;
MODEL Y1 -Y3 = A B A* B/SS3 NOUNI;
MANOVA H = _ALL_;

LSMEANS A* B/CL TDIFF PDIFF;
RUN;
PROC GLM;
CLASS A B;
MODEL Y1 -Y3 = A B/SS3 NOUNI;
MANOVA H = _ALL_;
LSMEANS A B/CL TDIFF PDIFF;
RUN;QUIT;
ODS HTML CLOSE;
```

【主要分析结果及解释】

以下是模型中含交互作用效应项的 GLM 过程输出的多元方差分析的结果：

Multivariate Analysis of Variance

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall A Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.40375481 | 6.89 | 3 | 14 | 0.0044 |

分析结果表明：A 对 Y1、Y2、Y3 整体的影响有统计学意义 (Wilks' $\lambda = 0.404$, $F = 6.89$, $P = 0.0044 < 0.05$)。

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall B Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.50025197 | 4.66 | 3 | 14 | 0.0185 |

分析结果表明：B 对 Y1、Y2、Y3 整体的影响有统计学意义 (Wilks' $\lambda = 0.500$, $F = 4.66$, $P = 0.0185 < 0.05$)。

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall A * B Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.80764324 | 1.11 | 3 | 14 | 0.3775 |

分析结果表明：A 与 B 的交互作用效应项对 Y1、Y2、Y3 整体的影响无统计学意义 (Wilks' $\lambda = 0.808$, $F = 1.11$, $P = 0.3775 > 0.05$)。此时，为了使分析结果更为准确，宜将模型中的交互作用效应项去掉，重新进行计算，重新估计主效应(由第二个 GLM 过程来实现)。

以下是模型中不含交互作用效应项的 GLM 过程输出的多元方差分析的结果。

Multivariate Analysis of Variance

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall A Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.40579476 | 7.32 | 3 | 15 | 0.0030 |

分析结果表明：A 对 Y1、Y2、Y3 整体的影响有统计学意义 (Wilks' $\lambda = 0.406$, $F = 7.32$, $P = 0.0030 < 0.05$)。

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall B Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.53049330 | 4.43 | 3 | 15 | 0.0203 |

分析结果表明：B 对 Y1、Y2、Y3 整体的影响有统计学意义 (Wilks' $\lambda = 0.530$, $F = 4.43$, $P = 0.0203 < 0.05$)。

专业结论：挤压度和添加剂剂量对塑料胶片的性能有影响，但挤压度和添加剂剂量之间不存在相互影响。

12.2.3 对例 12-3 资料进行含区组因素析因设计定量资料四元方差分析

【程序说明】 用“BLOCK”表示“区组因素”，用 A、B 分别表示两个实验因素，用“MD”、“LB”、“CA”和“AP”分别表示四个定量的观测指标。所需的 SAS 程序如下，程序名为 SAS-TJFX12_3.SAS。数据文件名为 DATA12_3.TXT，存放在 D 盘根目录下。

```
DATA SASTJFX12_3;
INFILE 'D:\DATA12_3.TXT';
INPUT BLOCK A B MD LB CA AP@@ ;
RUN;
ODS HTML FILE = "D:\OUTPUT12_3.HT-
ML";

PROC GLM;
CLASS BLOCK A B;
MODEL MD LB CA AP = BLOCK A B A* B/SS3
NOUNI;
MANOVA H = _ALL_;
LSMEANS A* B;
RUN;QUIT;
ODS HTML CLOSE;
```

【主要分析结果及解释】

Multivariate Analysis of Variance

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall BLOCK Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.45852358 | 1.64 | 20 | 123.66 | 0.0536 |

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall A Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.39150518 | 5.53 | 8 | 74 | <.0001 |

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall B Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.03833596 | 37.99 | 8 | 74 | <.0001 |

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall A * B Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.22853777 | 4.41 | 16 | 113.67 | <.0001 |

以上是多元方差分析的结果。区组因素“制模员”对 4 项指标整体的影响无统计学意义 ($F = 1.64$, $P = 0.0536 > 0.05$)，实验因素“消毒剂种类”和“消毒时间”对 4 项指标整体的影响均有统计学意义 ($F = 5.53$, $P < 0.0001$; $F = 37.99$, $P < 0.0001$)，两个实验因素的交互作用有统计学意义 ($F = 4.41$, $P < 0.0001$)。

The GLM Procedure

Least Squares Means

| A | B | MD LSMEAN | LB LSMEAN | CA LSMEAN | AP LSMEAN |
|---|---|------------|------------|------------|------------|
| 1 | 1 | 9.41300000 | 8.98400000 | 51.3786667 | 29.5338333 |
| 1 | 2 | 9.41750000 | 8.99283333 | 51.3941667 | 29.5580000 |
| 1 | 3 | 9.42400000 | 9.00933333 | 51.4088333 | 29.5463333 |
| 2 | 1 | 9.41733333 | 8.97733333 | 51.3801667 | 29.5278333 |
| 2 | 2 | 9.41666667 | 8.98000000 | 51.3896667 | 29.5430000 |
| 2 | 3 | 9.41516667 | 9.01116667 | 51.4120000 | 29.5480000 |
| 3 | 1 | 9.41500000 | 8.98250000 | 51.3755000 | 29.5328333 |
| 3 | 2 | 9.41616667 | 8.98516667 | 51.3846667 | 29.5340000 |
| 3 | 3 | 9.42100000 | 9.00550000 | 51.4060000 | 29.5468333 |

这是两个实验因素各水平组合下 4 项指标的最小二乘均数。

专业结论: 6 位印模试件制作技术员的制模精度很相近, “消毒剂种类”和“消毒时间”两个实验因素对消毒后模型的尺寸有影响, 而且两个实验因素之间存在相互影响。

12.2.4 对例 12-4 资料进行正交设计定量资料三元方差分析

【程序说明】所需的 SAS 程序如下, 程序名为 SASTJFX12_4.SAS。数据文件名为 DATA12_4.TXT, 存放在 D 盘根目录下。

```
DATA SASTJFX12_4;
INFILE 'D:\DATA12_4.TXT';
INPUT A F B C H E D G y1 y2 y3 @@ ;
RUN;
ODS HTML FILE = "D:\OUTPUT12_4.HT-
ML";
PROC GLM DATA = SASTJFX12_4;
CLASS A B C D E F G H;
MODEL y1 y2 y3 = A B C D E F G H A* F E*
H/SS3 NOUNI;

MANOVA H = _ALL_/PRINTE PRINTH;
MEANS A B C D E F G H;
RUN;
PROC GLM DATA = SASTJFX12_4;
CLASS A B C D E F G H;
MODEL y1 y2 y3 = D F/SS3 NOUNI;
MANOVA H = _ALL_/PRINTE PRINTH;
ODS HTML;
RUN;QUIT;
ODS HTML CLOSE;
```

注意: 本程序中没有体现效应项的筛选过程。实际应用中, 一般先筛掉 P 值较大的效应项, 在高阶交互效应项与低阶交互效应项对应 P 值接近的情况下, 优先筛掉高阶交互效应项。对本资料而言, 被筛掉的效应项从先到后依次为: E * H、A、B、E、H、G、A * F 和 C, 最后仅剩余 D 和 F 效应项对应的 P 值小于 0.05。

【主要分析结果及解释】

The GLM Procedure

Multivariate Analysis of Variance

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall A Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.60104086 | 0.19 | 6 | 4 | 0.9623 |

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall B Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.53625811 | 0.24 | 6 | 4 | 0.9387 |

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall C Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.09034413 | 1.55 | 6 | 4 | 0.3493 |

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall D Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.06922359 | 1.87 | 6 | 4 | 0.2840 |

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall E Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.43739518 | 0.34 | 6 | 4 | 0.8841 |

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall F Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.11331759 | 1.31 | 6 | 4 | 0.4133 |

MANOVA Test Criteria and F Approximations for the Hypothesis of No OverallG Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.22784234 | 0.73 | 6 | 4 | 0.6527 |

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall H Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.32960144 | 0.49 | 6 | 4 | 0.7897 |

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall A * F Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.14921060 | 1.06 | 6 | 4 | 0.5010 |

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall E * H Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.80220128 | 0.08 | 6 | 4 | 0.9958 |

以上是第一个过程步输出的多元方差分析的结果。对于观测指标 y1、y2 和 y3，各个因素对它们的整体影响及交互作用均无统计学意义。

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall D Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.36392774 | 4.38 | 6 | 40 | 0.0017 |

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall F Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.51477230 | 2.63 | 6 | 40 | 0.0306 |

以上是第二个过程步输出的多元方差分析的结果。对于观测指标 y1、y2 和 y3，因素 D、F 对它们的整体影响均有统计学意义 (Wilks' λ = 0.3639, F = 4.38, P = 0.0017 < 0.05; Wilks' λ = 0.5148, F = 2.63, P = 0.0306 < 0.05)，而其他因素或因素间交互效应项的影响无统计学意义。

专业结论：“碱洗搅拌速度”和“水洗温度”对反应航煤洁净度的“微水量”、“微量钠含量”

和“酸度”3 项观测指标有影响。不过,由于本研究为多元定量资料,比较复杂,已实施的实验也还不够完善,故很难在目前结合各因素各水平下各指标的均值情况以及航煤洁净度合格的条件,给出一个比较好的工艺条件。要达到此目标,尚需做进一步试验。

12.2.5 对例 12-5 资料进行具有一个重复测量的两因素设计定量资料二元方差分析

【程序说明】 下面是分析此资料的 SAS 程序,程序名为 SASTJFX12_5. SAS。程序中用“HEALTH”表示“受试对象患病情况”,用“REPETITION”表示受试对象编号,用“LF1、LF2、LF3”和“HF1、HF2、HF3”分别表示 2 个响应变量在 3 个不同时间点被重复观测的结果。数据文件名为 DATA12_5TXT,存放在 D 盘根目录下。语句“REPEATED RESPONSE 2 identity, TIME 3;”中的“RESPONSE 2 identity”表示有 2 个响应变量,这是较高级版本的 SAS 软件分析具有重复测量的多元定量资料的一种比较简便的方法,关键词“identity”一方面表示前面的“RESPONSE”为“响应变量”而非“重复测量因素”,另一方面表示要求分析过程中进行恒等变换(identity transformation),否则,将不能对资料进行正确分析;“TIME 3”表示“TIME”为重复测量因素,而且有 3 个水平。

```
DATA SASTJFX12_5;
INFILE 'D:\DATA12_5.TXT';
INPUT HEALTH $ REPETITION LF1 LF2
LF3 HF1 HF2 HF3 @@ ;
RUN;
ODS HTML FILE = "D:\OUTPUT12_5.HT-
ML";

PROC GLM DATA = SASTJFX12_5;
CLASS HEALTH;
MODEL LF1 LF2 LF3 HF1 HF2 HF3 =
HEALTH/NOUNI;
REPEATED RESPONSE2 identity, TIME 3;
LSMEANS HEALTH/CL;
RUN;QUIT;
ODS HTML CLOSE;
```

【主要分析结果及解释】

The GLM Procedure

Repeated Measures Analysis of Variance

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no RESPONSE Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.00861608 | 2128.65 | 2 | 37 | <.0001 |

这是对两个响应变量所对应的总体均值向量进行检验的多元方差分析结果: Wilks' $\lambda = 0.00861608$, $F = 2128.65$, $P < 0.0001$, 表明两个响应变量所对应的总体均值向量之间的差别有统计学意义。不过,这个结果对读者来说没有多大价值,可以不管它。

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no

RESPONSE * HEALTH Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.68646392 | 8.45 | 2 | 37 | 0.0009 |

这是对主效应 HEALTH 检验的多元方差分析结果: Wilks' $\lambda = 0.68646392$, $F = 8.45$, $P = 0.0009$, 表明 HEALTH 对两个响应变量整体的影响具有统计学意义。

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of

no RESPONSE * TIME Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.71913580 | 3.42 | 4 | 35 | 0.0185 |

这是对主效应 TIME 检验的多元方差分析结果: Wilks' $\lambda = 0.71913580$, $F = 3.42$, $P = 0.0185$, 表明 TIME 对两个响应变量整体的影响具有统计学意义。

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no
RESPONSE * TIME * HEALTH Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.84002722 | 1.67 | 4 | 35 | 0.1798 |

这是对交互效应 TIME * HEALTH 检验的多元方差分析结果: Wilks' $\lambda = 0.84002722$, $F = 1.67$, $P = 0.1798 > 0.05$, 表明交互效应 TIME * HEALTH 对两个响应变量整体的影响无统计学意义。

Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|--------|----|-------------|-------------|---------|--------|
| HEALTH | 1 | 22.28332042 | 22.28332042 | 8.48 | 0.0060 |
| Error | 38 | 99.81806917 | 2.62679129 | | |

这是对主效应 HEALTH 检验的一元方差分析结果(先将各指标在各个时间点的测量结果通过某种方式综合成一个单一的指标): $F = 8.48$, $P = 0.0060$, 表明 HEALTH 对两个响应变量的影响有统计学意义。当一元方差分析的结果与前面给出的多元方差分析的结果矛盾时,应以多元方差分析的结果为准,因为多元方差分析的结果一般较为可靠。

Least Squares Means

| HEALTH | LF1 LSMEAN | LF2 LSMEAN | LF3 LSMEAN | HF1 LSMEAN | HF2 LSMEAN | HF3 LSMEAN |
|--------|------------|------------|------------|------------|------------|------------|
| HA | 4.85850000 | 4.89200000 | 4.90900000 | 3.58800000 | 3.75700000 | 3.82100000 |
| SAP | 5.24450000 | 5.69500000 | 5.73150000 | 3.97850000 | 4.39050000 | 4.44200000 |

这里列出了两组各时间点上响应变量的最小二乘均值。

专业结论: 稳定型心绞痛患者的低、高频心电频谱值均高于正常人的低、高频心电频谱值。不管是正常人还是稳定型心绞痛患者, 他们的低、高频心电频谱值都随观察时间的推移而增大。

12.2.6 对例 12-6 资料进行两因素析因设计五元定量资料的二元协方差分析

【程序说明】 下面是分析此资料的 SAS 程序, 程序名为 SASTJFX12_6.SAS。此程序的第一个 GLM 过程步主要用于考察协变量与分组变量之间的交互作用有无统计学意义。第二个 GLM 过程步用于完成协方差分析。数据文件名为 DATA12_6.TXT, 存放在 D 盘根目录下。

```
DATA SASTJFX12_6;
INFILE 'D:\DATA12_6.TXT';
INPUT ID WORK AGE WORK_TIME SMOKING
P21 P53 SCE NUM_OF_ABERRATION NUM
OF_CELLS @@ ;
RUN;
ODS HTML FILE = "D:\OUTPUT12_6.HTML";

MANOVA H = _ALL_;
RUN;
PROC GLM;
CLASS WORK SMOKING;
MODEL P21 P53 SCE NUM_OF_ABERRATION
NUM_OF_CELLS =
AGE WORK_TIME WORK SMOKING
```

```

PROC GLM;
CLASS WORK SMOKING;
MODEL P21 P53 SCE NUM_OF_ABERRATION
NUM_OF_CELLS =
AGE WORK _TIME WORK SMOKING AGE *
SMOKING AGE * WORK WORK _TIME * SMOK-
ING WORK _TIME * WORK WORK * SMOKING
/SS3 NOUNI;
WORK* SMOKING/SS3 NOUNI;
MANOVA H = _ALL_;
RUN;QUIT;
ODS HTML CLOSE;

```

【主要分析结果及解释】

以下是第一个过程步输出的统计分析结果。此过程步主要用来考察协方差分析应用的前提条件之一：各组由自变量推测响应变量的回归方程的总体回归系数相等。所以，其他对于主效应的分析可不必在意，仅注意协变量与其他自变量的交互效应有无统计学意义即可。

Multivariate Analysis of Variance

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall AGE Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.66822254 | 0.89 | 5 | 9 | 0.5240 |

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall WORK_TIME Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.68550502 | 0.83 | 5 | 9 | 0.5617 |

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall WORK Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.83100059 | 0.37 | 5 | 9 | 0.8596 |

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall SMOKING Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.44664173 | 0.89 | 10 | 18 | 0.5571 |

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall AGE * SMOKING Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.42023678 | 0.98 | 10 | 18 | 0.4951 |

分析结果表明：AGE * SMOKING 对五个响应变量整体的影响无统计学意义 (Wilks' $\lambda = 0.4202$, $F = 0.98$, $P = 0.4951 > 0.05$)。

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall AGE * WORK Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.71324877 | 0.72 | 5 | 9 | 0.6225 |

分析结果表明：AGE * WORK 对五个响应变量整体的影响无统计学意义 (Wilks' $\lambda = 0.7132$, $F = 0.72$, $P = 0.6225 > 0.05$)。

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall WORK_TIME * SMOKING Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.42332855 | 0.97 | 10 | 18 | 0.5025 |

分析结果表明: WORK_TIME * SMOKING 对五个响应变量整体的影响无统计学意义 (Wilks' $\lambda = 0.4233$, $F = 0.97$, $P = 0.5025 > 0.05$)。

MANOVA Test Criteria and Exact F Statistics for the
Hypothesis of No Overall WORK_TIME * WORK Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.60367936 | 1.18 | 5 | 9 | 0.3890 |

分析结果表明: WORK_TIME * WORK 对五个响应变量整体的影响无统计学意义 (Wilks' $\lambda = 0.6037$, $F = 1.18$, $P = 0.3890 > 0.05$)。

以上几项协变量与自变量的交互效应均无统计学意义,说明本资料满足协方差分析的应用条件。

以下为第二个过程步产生的分析结果,即协方差分析的最终结果。

Multivariate Analysis of Variance

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall AGE Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.82540596 | 0.63 | 5 | 15 | 0.6767 |

分析结果表明: AGE 对五个响应变量整体的影响无统计学意义 (Wilks' $\lambda = 0.8254$, $F = 0.63$, $P = 0.6767 > 0.05$)。

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall WORK_TIME Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.91687311 | 0.27 | 5 | 15 | 0.9214 |

分析结果表明: WORK_TIME 对五个响应变量整体的影响无统计学意义 (Wilks' $\lambda = 0.9169$, $F = 0.27$, $P = 0.9214 > 0.05$)。

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall WORK Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.82344203 | 0.64 | 5 | 15 | 0.6707 |

分析结果表明: WORK 对五个响应变量整体的影响无统计学意义 (Wilks' $\lambda = 0.8234$, $F = 0.64$, $P = 0.6707 > 0.05$)。

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall SMOKING Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.82367806 | 0.31 | 10 | 30 | 0.9740 |

分析结果表明: SMOKING 对五个响应变量整体的影响无统计学意义 (Wilks' $\lambda = 0.8237$, $F = 0.31$, $P = 0.974 > 0.05$)。

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall WORK * SMOKING Effect

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.47175281 | 1.37 | 10 | 30 | 0.2421 |

分析结果表明: WORK * SMOKING 对五个响应变量整体的影响无统计学意义 (Wilks' $\lambda = 0.4718$, $F = 1.37$, $P = 0.2421 > 0.05$)。

专业结论：就本例资料而言，工种、年龄、工龄和吸烟量及工种与吸烟量的交互项对工人各项生物标志物水平均无影响。

不过，需要说明的是，本例资料是假设的，此处仅起“演示用 SAS 实现析因设计多元定量资料多元协方差分析”之作用，并不代表实际问题的真实情况。

12.3 本章小结

本章结合实例介绍了用 SAS 实现随机区组设计(也称配伍组设计)、析因设计、含区组因素的析因设计、正交设计和重复测量设计多元定量资料的多元方差分析方法及析因设计多元定量资料的一元或多元协方差分析方法。本章的所有例题均给出了 SAS 程序、分析结果、统计学结论和专业结论，读者在解决类似的实际问题时完全可以如法炮制。当然，要使资料的分析正确，读者还必须注意以下几点：(1)检查试验设计或调查设计的正确性、合理性；(2)检查资料是否为多元定量资料，是否需要做多元统计分析；(3)检查多元定量资料是否满足多元方差分析或一元、多元协方差分析的前提条件。有关多元方差分析或一元、多元协方差分析的前提条件的检查方法比较复杂，本章未予介绍，请读者查阅相关文献。

(周诗国)

第 4 篇 对定性结果进行差异性分析

第 13 章 单因素设计一元定性资料差异性分析

本章主要内容是用 SAS 软件实现单因素设计一元定性资料的统计分析。与单因素设计一元定量资料类似，该类资料从设计类型上分为单组设计、配对设计、成组设计、单因素多水平设计。本章结合具体实例，阐述其相应的统计分析过程。

13.1 单组设计一维表资料统计分析

13.1.1 问题与数据

【例 13-1】 为了调查某工厂产品生产合格率的情况，随机抽取 1000 件该厂产品进行检验，发现其中 978 件合格，22 件不合格。已知业内相同产品的生产合格率为 98%，问：该工厂产品合格率是否低于业内平均水平？

13.1.2 对数据结构的分析

该资料中所有观测对象未按其他因素分组，它们均处在同一个组中，而观测指标是一个二值变量（即“合格与否”）。因此从设计类型角度来看，它属于单组设计一元定性资料。

13.1.3 分析目的与统计分析方法的选择

该研究的目的是要考察该 1000 件产品所代表的总体与业内相同产品所代表的总体，产品合格率差别是否有统计学意义，由于结果变量是二值变量，所以可以用二项检验的方法对数据进行统计处理。

13.1.4 SAS 程序中重要内容的说明

分析此资料所需的 SAS 程序（程序名为 SASTJFX13_1.SAS）如下：

| | |
|--|--|
| <pre>DATA SASTJFX13_1; INPUT group count; CARDS; 1 978 2 22 ; RUN;</pre> | <pre>PROC FREQ; TABLES group/binomial (p=0.98); WEIGHT count; RUN;</pre> |
|--|--|

数据步, 建立名为 pgm013_1 的数据集, group 是分组变量, count 是频数变量。过程步, 调用 FREQ 过程, 用 TABLES 语句加变量 group 表示绘制一维列联表; binomial 选项表示按照二项分布原理对统计量进行计算和检验, 用($p=0.98$)指定标准率的大小, 最后用 WEIGHT 语句指定频数变量 count。

13.1.5 主要分析结果及解释

group 的二项分布比例 = 1

| | |
|----------|--------|
| 比例 | 0.9780 |
| 渐近标准误差 | 0.0046 |
| 95% 置信下限 | 0.9689 |
| 95% 置信上限 | 0.9871 |

以上给出了 group = 1, 即 1000 件产品所代表总体的产品合格率估计值及其 95% 置信区间, 合格率为 0.9780, 置信区间为(0.9689, 0.9871)。

H0 检验: 比例 = 0.98

| | |
|-------------|---------|
| H0 下的渐近标准误差 | 0.0044 |
| Z | -0.4518 |
| 单侧 Pr < Z | 0.3257 |
| 双侧 Pr > Z | 0.6514 |

以上是进行二项分布检验所得到的结果, 其原假设 H0 是“1000 件产品所代表的总体与业内相同产品代表的总体, 产品合格率无差别”。

根据二项检验结果, $Z = -0.4518$, $P = 0.6514 > 0.05$, 在检验水准 $\alpha = 0.05$ 条件下, 不拒绝原假设, 即还不能认为两总体的产品合格率差异有统计学意义。因此, 可以认为该工厂的产品合格率与业内平均水平基本相同。

13.2 配对设计四格表资料统计分析

13.2.1 问题与数据

【例 13-2】 两家评审机构(A 和 B)对同一组评价项目的评审结果如表 13-1 所示。问: 两家评审机构的评价结果是否一致?

表 13-1 两家评审机构对同一组评价项目的评审结果

| 机构 A 评审结果 | 项目数 | | | |
|--------------|------------|----|-----|----|
| | 机构 B 评审结果: | 通过 | 未通过 | 合计 |
| 通过 | | 55 | 12 | 67 |
| 未通过 | | 8 | 17 | 25 |
| 合计 | | 63 | 29 | 92 |

13.2.2 对数据结构的分析

该资料是配对设计四格表资料, 即同一项目均由两机构来评价, 且评价结果是相互对立的两种结果。整理成这种配对四格表形式, 可以直观地看出评价结果一致和不一致的频数分布情况。

13.2.3 分析目的与统计分析方法的选择

比较两个机构评审结果不一致部分的差别是否有统计学意义, 可采用 McNemar χ^2 检验; 比较两个机构评审结果的观测一致率与期望一致率之间的差别是否有统计学意义可采用 Kappa 检验。

13.2.4 SAS 程序中重要内容的说明

分析此资料所需的 SAS 程序(程序名为 SASTJFX13_2.SAS)如下：

```
DATA SASTJFX13_2;
  DO A = 1 TO 2;
    DO B = 1 TO 2;
      INPUT F @@ ;
      OUTPUT;
    END;
  END;
CARDS;
55 12
8 17
;
RUN;
```

```
PROC FREQ;
  WEIGHT F;
  TABLES A* B/AGREE;
  TEST KAPPA;
RUN;
```

数据步，建立名为 pgm013_2 的数据集，通过变量 A、B、F，分别读入行号、列号、每个基本格子中的实际频数。过程步，调用 FREQ 过程，指定频数变量 F，用 TABLES A * B 语句表示绘制二维列联表；AGREE 语句是配对四格表输出 McNemar χ^2 统计量的关键语句；TEST KAPPA 用于对 Kappa 系数作假设检验。

13.2.5 主要分析结果及解释

| McNemar 检验 | |
|------------|--------|
| 统计量(S) | 0.8000 |
| 自由度 | 1 |
| Pr > S | 0.3711 |

以上是 McNemar χ^2 检验的输出结果，此处使用的是未进行校正的计算公式。S = 0.8000，P = 0.3711。

| 简单 Kappa 系数 | |
|-------------|--------|
| Kappa | 0.4770 |
| 渐近标准误差 | 0.0999 |
| 95% 置信下限 | 0.2812 |
| 95% 置信上限 | 0.6728 |

以上给出了 Kappa 值及其渐近标准误、95% 置信区间的上限和下限。

| H0 检验：Kappa = 0 | |
|-----------------|---------|
| H0 下的渐近标准误差 | 0.1037 |
| Z | 4.6002 |
| 单侧 Pr > Z | < .0001 |
| 双侧 Pr > Z | < .0001 |

以上部分是对原假设“Kappa = 0”的假设检验结果，Z = 4.6002，P < 0.0001。

经 McNemar χ^2 检验，两机构评价结果不一致部分的差异没有统计学意义，说明两机构评价结果的不一致部分差异很小。经 Kappa 检验，两个机构评审结果的观测一致率与期望一致率之间的差异有统计学意义，说明两机构的评价结果具有一致性。但 Kappa 系数的具体取值

为 0.4770, 其 95% 置信区间为(0.2812, 0.6728), 可以认为 Kappa 值尚不够大, 说明两机构的一致性仅在统计学上有一定意义, 但实际价值并不高(可根据专业知识对预期的 Kappa 值进行界定)。

13.3 配对设计扩大形式的方表资料统计分析

13.3.1 问题与数据

【例 13-3】用快速法和 ELISA 法对同一批样品进行抗体检测试验, 结果见表 13-2。问: 这两种检测方法所得结果是否一致?

表 13-2 两种测定方法同时检测抗体得 73 个样品的检测结果

| 快速法 检测结果 | 样品数 | | | | | 合计 |
|-------------|----------|----|----|----|-----|----|
| | ELISA 法: | - | + | ++ | +++ | |
| - | | 15 | 0 | 2 | 3 | 20 |
| + | | 2 | 19 | 1 | 2 | 24 |
| ++ | | 1 | 3 | 17 | 0 | 21 |
| +++ | | 0 | 2 | 0 | 6 | 8 |
| 合计 | | 18 | 24 | 20 | 11 | 73 |

13.3.2 对数据结构的分析

该资料行变量和列变量性质相同且取值水平和含义也相同, 它实际上是配对四格表资料的扩大, 设计上属于配对设计扩大形式的方形列联表资料, 列联表分类上属于双向有序且属性相同的 $R \times C$ 列联表。

13.3.3 分析目的与统计分析方法的选择

该例主要目的是解决两种检测方法结果是否一致的问题, 常用的统计分析方法是一致性检验, 即 Kappa 检验。

13.3.4 SAS 程序中重要内容的说明

分析此资料所需的 SAS 程序如下, 程序名为 SASTJFX13_3.SAS。

```

DATA SASTJFX13_3;
  DO A = 1 TO 4;
    DO B = 1 TO 4;
      INPUT F @@ ;
      OUTPUT;
    END;
  END;
CARDS;
15  0  2  3
 2 19  1  2
 1  3 17  0
 0  2  0  6
;
RUN;

PROC FREQ;
  WEIGHT F;
  TABLES A* B;
  TEST KAPPA;
RUN;

```

程序与上一例类似，此处 TEST 语句中的 KAPPA 可以用 AGREE 代替，输出结果比用 KAPPA 时多了对加权 KAPPA 进行假设检验的结果；若没有“TEST KAPPA;”语句，仅在“TABLES A * B;”语句后加上 AGREE 选项，即写成“TABLES A * B/AGREE;”，此时，仅输出关于简单和加权 KAPPA 统计量值、渐近标准误和 95% 置信区间的下限和上限，不对 KAPPA 系数进行假设检验；若将现在的 TEST 语句改成“TEST AGREE;”，可同时输出关于简单和加权 KAPPA 的统计量值、渐近标准误、95% 置信区间以及假设检验结果；若将 TEST 语句改成“TEST WTKAP;”则可输出关于简单和加权 KAPPA 的统计量值、渐近标准误、95% 置信区间，仅给出对加权 KAPPA 检验的结果。

13.3.5 主要分析结果及解释

| 简单 Kappa 系数 | |
|-------------|--------|
| Kappa | 0.6994 |
| 渐近标准误差 | 0.0658 |
| 95% 置信下限 | 0.5704 |
| 95% 置信上限 | 0.8283 |

以上给出了简单 Kappa 统计量值及其 95% 置信区间。

| H0 检验: Kappa = 0 | |
|------------------|---------|
| H0 下的渐近标准误差 | 0.0699 |
| Z | 10.0044 |
| 单侧 Pr > Z | < .0001 |
| 双侧 Pr > Z | < .0001 |

以上部分是对原假设“Kappa = 0”的假设检验结果，Z = 10.0044，P < 0.0001。

经 Kappa 检验(或称一致性检验)可知，两种检测方法检测结果的观测一致率与期望一致率之间的差别有统计学意义，说明两种方法检测结果具有一致性。结合 Kappa 值的 95% 置信区间(0.5704, 0.8283)，可以认为 Kappa 比较接近于 1，说明两检测方法的一致性具有较高的实际意义。

13.4 成组设计横断面研究四格表资料统计分析

13.4.1 问题与数据

【例 13-4】 某研究随机抽取了某大学四年级学生 124 人，调查大学英语六级通过情况，结果见表 13-3。问：该大学男生和女生英语六级通过率有无差别？

表 13-3 某大学四年级男、女生大学英语六级通过情况

| 性别 | 人 数 | | | |
|----|-----------|----|-----|-----|
| | 英语六级通过情况： | 通过 | 未通过 | 合计 |
| 男 | | 41 | 32 | 73 |
| 女 | | 43 | 8 | 51 |
| 合计 | | 84 | 40 | 124 |

13.4.2 对数据结构的分析

该资料设计上属于结果变量为二值变量的成组设计定性资料，列联表分类上属于横断面研究设计四格表资料。

13.4.3 分析目的与统计分析方法的选择

该例目的是比较两个性别组英语六级通过率是否相同,可采用一般 χ^2 检验或 Fisher 精确检验来处理。

13.4.4 SAS 程序中重要内容的说明

分析此资料所需的 SAS 程序如下,程序名为 SASTJFX13_4.SAS。

| | |
|--|--|
| <pre>DATA SASTJFX13_4; DO A=1 TO 2; DO B=1 TO 2; INPUT F @@ ; OUTPUT; END; END; CARDS; 41 32 43 8 ; RUN;</pre> | <pre>PROC FREQ; WEIGHT F; TABLES A*B/CHISQ; RUN;</pre> |
|--|--|

该程序关键选项是 TABLES 语句中的 CHISQ,目的是输出几种常用的 χ^2 检验统计量,列联表为四格表时,还可以输出 Fisher 精确检验结果。

13.4.5 主要分析结果及解释

| 统计量 | 自由度 | 值 | 概率 |
|--------------------|-----|---------|--------|
| 卡方 | 1 | 10.8871 | 0.0010 |
| 似然比卡方 | 1 | 11.5432 | 0.0007 |
| 连续校正卡方 | 1 | 9.6370 | 0.0019 |
| Mantel-Haenszel 卡方 | 1 | 10.7993 | 0.0010 |

以上给出了若干 χ^2 检验统计量,通常查看一般 χ^2 检验结果即可,当总频数 $n > 40$ 但至少有一个理论频数 $1 < T \leq 5$ 时,可选用连续校正的 χ^2 检验。

Fisher 精确检验

| | |
|----------------|-----------|
| 单元格(1,1)频数(F) | 41 |
| 左侧 Pr \leq F | 7.396E-04 |
| 右侧 Pr \geq F | 0.9998 |
| 表概率(P) | 5.806E-04 |
| 双侧 Pr \leq P | 9.474E-04 |

以上是 Fisher 精确检验结果,本例 $P = 9.474E-04$,结论与基于上面 χ^2 检验结果做出的结论一致。当总频数 $n < 40$ 或有理论频数 $T < 1$ 时(SAS 输出结果会有提示),必须选用 Fisher 精确检验。

本例符合一般 χ^2 检验的使用条件,因此只需要查看一般 χ^2 检验结果即可: $\chi^2 = 10.8871$, $P = 0.001$,按照 0.05 的检验水准,拒绝原假设,接受备择假设,可以认为男、女生大学英语六级通过率不同,结果表明女生的通过率高于男生。

13.5 成组设计队列研究四格表资料统计分析

13.5.1 问题与数据

【例 13-5】 某研究者随机选取了 565 名调查对象，其中血压偏高者 80 名，血压正常者 485 名，经过多年追踪观察得到如表 13-4 所示的资料。试分析当初血压情况对调查对象后来是否患冠心病可能造成什么样的影响。

表 13-4 血压与冠心病关系队列研究结果

| 血压 情况 | 例 数 | | | |
|----------|--------|----|-----|-----|
| | 冠心病情况： | 患病 | 未患病 | 合计 |
| 偏高 | | 19 | 61 | 80 |
| 正常 | | 20 | 465 | 485 |
| 合计 | | 39 | 526 | 565 |

13.5.2 对数据结构的分析

该资料是成组设计队列研究四格表资料。队列研究设计是通过在不同暴露水平的对象进行追踪观察，随访观察疾病发生情况，从而判断该因素与发病之间有无关联。

13.5.3 分析目的与统计分析方法的选择

该例研究目的是比较两个血压组冠心病发病的概率是否相同，可以在一般 χ^2 检验的基础上，采用 Mantel-Haenszel χ^2 对 RR 值(相对危险度)进行检验。

13.5.4 SAS 程序中重要内容的说明

分析此资料所需的 SAS 程序如下，程序名为 SASTJFX13_5.SAS。

```
DATA SASTJFX13_5;
  DO A=1 TO 2;
    DO B=1 TO 2;
      INPUT F @@ ; OUTPUT;
    END;
  END;
CARDS;
19 61
20 465
;
RUN;
```

```
PROC FREQ;
  WEIGHT F;
  TABLES A* B/CHISQ CMH RISKDIFF
  alpha=0.05;
RUN;
```

该程序在 CHISQ 后面加 CMH 选项，输出 Cochran-Mantel-Haenszel 统计量，列联表为 2×2 表时，它可以输出 OR 和 RR 估计值及置信区间。RISKDIFF 选项可以输出两个血压组代表的总体冠心病患病率之差的置信区间。“alpha=0.05”指定求出 95% 置信区间。

13.5.5 主要分析结果及解释

| 统计量 | 自由度 | 值 | 概率 |
|--------------------|-----|---------|--------|
| 卡方 | 1 | 41.1629 | <.0001 |
| 似然比卡方 | 1 | 29.3491 | <.0001 |
| 连续校正卡方 | 1 | 38.1655 | <.0001 |
| Mantel-Haenszel 卡方 | 1 | 41.0901 | <.0001 |

这是 χ^2 检验的结果, 其中一般 χ^2 检验中 $\chi^2 = 41.1629$, $P < 0.0001$ 。

第 1 列风险估计

| | 风险 | 渐近标准误差 | (渐近的)95% 置信限 | | (精确的)95% 置信限 | |
|-------|--------|--------|-----------------|--------|-----------------|--------|
| 第 1 行 | 0.2375 | 0.0476 | 0.1442 | 0.3308 | 0.1495 | 0.3458 |
| 第 2 行 | 0.0412 | 0.0090 | 0.0235 | 0.0589 | 0.0254 | 0.0630 |
| 合计 | 0.0690 | 0.0107 | 0.0481 | 0.0899 | 0.0495 | 0.0932 |
| 差值 | 0.1963 | 0.0484 | 0.1013 | 0.2912 | | |

差值为(行 1 ~ 行 2)

这是血压偏高组、血压正常组、两组合计的冠心病患病率估计值及其 95% 置信区间, 此外还包括两组患病率的差值及其 95% 置信区间。

Cochran-Mantel-Haenszel 统计量(基于表得分)

| 统计量 | 对立假设 | 自由度 | 值 | 概率 |
|-----|---------|-----|---------|--------|
| 1 | 非零相关 | 1 | 41.0901 | <.0001 |
| 2 | 行均值得分差值 | 1 | 41.0901 | <.0001 |
| 3 | 一般关联 | 1 | 41.0901 | <.0001 |

以上是 3 种不同检验方法对 RR 值的检验结果, 统计量均为 41.0901, PRR 值与 1 之间的差异具有统计学意义。

普通相对风险的估计值(行 1/行 2)

| 研究类型 | 方法 | 值 | 95% 置信限 | |
|---------------------|-----------------|--------|---------|---------|
| 案例对照 (优比) | Mantel-Haenszel | 7.2418 | 3.6605 | 14.3269 |
| | Logit | 7.2418 | 3.6605 | 14.3269 |
| Cohort (第 1 列风险) | Mantel-Haenszel | 5.7594 | 3.2193 | 10.3034 |
| | Logit | 5.7594 | 3.2193 | 10.3034 |
| Cohort (第 2 列风险) | Mantel-Haenszel | 0.7953 | 0.7028 | 0.9000 |
| | Logit | 0.7953 | 0.7028 | 0.9000 |

以上是 OR 和 RR 估计值及其置信区间, 本资料属于队列研究资料, 因此需要看“Cohort”对应的 RR 值, 第 1 列(患冠心病)的危险度 $RR = 5.7594$, 其 95% 置信区间为(3.2193, 10.3034)。

根据一般 χ^2 检验结果, 按照 0.05 的检验水准, 拒绝原假设, 接受备择假设, 可以认为血压偏高者和血压正常者冠心病患病率不同; 根据 RR 值及其置信区间, 可以认为若干年后患冠心病的风险, 血压偏高者是血压正常者的 3 ~ 10 倍。高血压组代表的总体比正常血压组代表的总体的冠心病患病率高 0.1963, 其 95% 置信区间为(0.1013, 0.2912)。

13.6 成组设计病例对照研究四格表资料统计分析

13.6.1 问题与数据

【例 13-6】在乳牙龋齿与喂养方式的关系研究中, 研究者对 3 ~ 4 岁儿童进行了病例对照研究, 分别选取患龋儿童 103 人和未患龋儿童 157 人, 调查出生后 6 个月喂养方式是否相同, 调查结果见表 13-5。

13.6.2 对数据结构的分析

该资料是成组设计病例对照研究四格表资料。病例对照设计是以确诊的患者作为病例，以不患该病但具有可比性的个体作为对照，收集既往危险因素的暴露史，用统计学方法比较两组中危险因素的暴露比例，从而判定因素与疾病之间是否存在统计学关联。

表 13-5 喂养方式与乳牙龋齿关系病例对照研究结果

| 喂养方式 | 例 数 | | |
|-------|-------|-----|-----|
| | 乳牙龋齿： | 患龋 | 未患龋 |
| 母乳 | | 37 | 81 |
| 人工或混合 | | 66 | 76 |
| 合计 | | 103 | 157 |

13.6.3 分析目的与统计分析方法的选择

该例研究目的是比较病例和对照组中危险因素的暴露比例是否相同，可以在一般 χ^2 检验的基础上，采用 Mantel-Haenszel χ^2 对 OR 值(优势比)进行检验。

13.6.4 SAS 程序中重要内容的说明

SAS 程序如下，名为 SASTJFX13_6.SAS。

```
DATA SASTJFX13_6;
  DO A=1 TO 2;
    DO B=1 TO 2;
      INPUT F @@ ; OUTPUT;
    END;
  END;
CARDS;
37 81
66 76
;
RUN;
```

```
PROC FREQ;
  WEIGHT F;
  TABLES A* B/CHISQ CMH;
RUN;
```

程序与上一例基本相同，用 CMH 选项输出 Cochran-Mantel-Haenszel 统计量。

13.6.5 主要分析结果及解释

| 统计量 | 自由度 | 值 | 概率 |
|--------------------|-----|--------|--------|
| 卡方 | 1 | 6.1614 | 0.0131 |
| 似然比卡方 | 1 | 6.2172 | 0.0127 |
| 连续校正卡方 | 1 | 5.5454 | 0.0185 |
| Mantel-Haenszel 卡方 | 1 | 6.1377 | 0.0132 |

这是 χ^2 检验的结果，其中一般 χ^2 检验中 $\chi^2=6.1614$ ， $P=0.0131$ 。

Cochran-Mantel-Haenszel 统计量(基于表得分)

| 统计量 | 对立假设 | 自由度 | 值 | 概率 |
|-----|---------|-----|--------|--------|
| 1 | 非零相关 | 1 | 6.1377 | 0.0132 |
| 2 | 行均值得分差值 | 1 | 6.1377 | 0.0132 |
| 3 | 一般关联 | 1 | 6.1377 | 0.0132 |

以上是 3 种不同检验方法对 OR 值的检验结果，统计量均为 6.1377， $P=0.0132<0.05$ ，可认为在总体上 OR 值与 1 之间的差异具有统计学意义。

普通相对风险的估计值(行 1/行 2)

| 研究类型 | 方法 | 值 | 95% 置信限 | |
|---------------------|-----------------|--------|---------|--------|
| 案例对照 (优比) | Mantel-Haenszel | 0.5260 | 0.3159 | 0.8759 |
| | Logit | 0.5260 | 0.3159 | 0.8759 |
| Cohort (第 1 列风险) | Mantel-Haenszel | 0.6746 | 0.4899 | 0.9291 |
| | Logit | 0.6746 | 0.4899 | 0.9291 |
| Cohort (第 2 列风险) | Mantel-Haenszel | 1.2826 | 1.0544 | 1.5601 |
| | Logit | 1.2826 | 1.0544 | 1.5601 |

以上是 OR 和 RR 估计值及其置信区间, 本资料属于病例对照研究资料, 因此需要看“病例对照”对应的 OR 值, 结果显示: $OR = 0.5260$, 其 95% 置信区间为 $(0.3159, 0.8759)$ 。注意, 此区间未包含 1。

根据一般 χ^2 检验结果, 按照 0.05 的检验水准, 拒绝原假设, 接受备择假设, 可以认为喂养方式不同乳牙龋齿患病率不同; 根据 OR 值及其置信区间, 可以认为: 母乳喂养与人工或混合喂养相比, 对乳牙龋齿具有一定的保护作用。也就是说, 人工喂养或混合喂养比母乳喂养易于导致乳牙龋齿。

13.7 成组设计结果变量为多值有序变量的 $2 \times C$ 表资料统计分析

13.7.1 问题与数据

【例 13-7】 为了调查北京和天津两地的居住环境, 随机抽取了两地居民各 120 人, 分别调查他们对居住环境的满意程度, 结果见表 13-6。试比较北京和天津两地居民的居住环境满意程度。

表 13-6 北京、天津两地居住环境满意程度比较

| 城市 | 人 数 | | | |
|----|-------|----|-----|-----|
| | 满意程度: | 满意 | 一般 | 不满意 |
| 北京 | | 45 | 48 | 27 |
| 天津 | | 29 | 65 | 26 |
| 合计 | | 74 | 113 | 53 |

13.7.2 对数据结构的分析

本例结果变量为多值有序变量, 原因变量为二值变量, 因此该资料是成组设计多值有序 $2 \times C$ 表资料。

13.7.3 分析目的与统计分析方法的选择

结果变量为多值有序变量, 可采用秩和检验方法来处理。

13.7.4 SAS 程序中重要内容的说明

分析此资料所需的 SAS 程序如下, 程序名为 SASTJFX13_7.SAS。

```
DATA SASTJFX13_7;
  DO A=1 TO 2;
    DO B=1 TO 3;
      INPUT F @@ ; OUTPUT;
    END;
  END;
CARDS;
45 48 27
29 65 26
;
```

```
PROC NPAR1WAY WILCOXON;
  CLASS A; VAR B; FREQ F;
RUN;
```

数据步，建立名为 pgm013_8 的数据集，变量 A、B、F 分别表示行号、列号、每个基本格子上的实际频数；过程步，调用单因素非参数检验过程 NPAR1WAY，WILCOXON 语句要求进行秩和检验：两组比较时进行 Wilcoxon 秩和检验及 Kruskal-Wallis 检验，多组比较时进行 Kruskal-Wallis 检验；最后用 CLASS 指定分类变量，VAR 指定观测变量，FREQ 指定频数变量。

13.7.5 主要分析结果及解释

| Wilcoxon Two-Sample Test | |
|--------------------------|------------|
| Statistic | 13753.5000 |
| Normal Approximation | |
| Z | -1.4193 |
| One-Sided Pr < Z | 0.0779 |
| Two-Sided Pr > Z | 0.1558 |

以上是对两组居住环境满意程度进行 Wilcoxon 秩和检验的结果， $Z = -1.4193$ ， $P = 0.1558 > 0.05$ 。

根据对两组进行秩和检验的结果， $P > 0.05$ ，不拒绝原假设，根据现有资料，尚不能认为两地居民对居住环境的满意程度有明显差别。

13.8 成组设计结果变量为多值名义变量的 $2 \times C$ 表资料统计分析

13.8.1 问题与数据

【例 13-8】 某研究调查汉族和回族 ABO 血型分布情况，分别随机抽取宁夏某大学汉族和回族学生各 500 人，所得结果见表 13-7。试比较汉族和回族在血型分布上的差别有无统计学意义。

表 13-7 汉族和回族 ABO 血型分布情况比较

| 民族 | 人 数 | | | | | |
|----|---------|-----|-----|-----|-----|------|
| | ABO 血型： | A | B | O | AB | 合计 |
| 汉族 | | 135 | 146 | 159 | 60 | 500 |
| 回族 | | 147 | 155 | 129 | 69 | 500 |
| 合计 | | 282 | 301 | 288 | 129 | 1000 |

13.8.2 对数据结构的分析

本例结果变量为多值名义变量，原因变量为二值变量，因此该资料是成组设计多值名义 $2 \times C$ 表资料。

13.8.3 分析目的与统计分析方法的选择

该例目的是比较两个民族组 ABO 血型频数分布是否相同，可采用一般 χ^2 检验或 Fisher 精确检验来处理。

13.8.4 SAS 程序中重要内容的说明

分析此资料所需的 SAS 程序如下，程序名为 SASTJFX13_8.SAS。

```

DATA SASTJFX13_8;
  DO A=1 TO 2;
    DO B=1 TO 4;
      INPUT F @@ ; OUTPUT;
    END;
  END;
CARDS;
135 146 159 60
147 155 129 69
;
RUN;

PROC FREQ;
  WEIGHT F;
  TABLES A* B/CHISQ;
  ExactFisher;
RUN;

```

由于本例资料不是 2×2 表资料，故程序中须输入 Exact Fisher 要求系统输出 Fisher 精确检验的结果。

13.8.5 主要分析结果及解释

| 统计量 | 自由度 | 值 | 概率 |
|--------------------|-----|--------|--------|
| 卡方 | 3 | 4.5326 | 0.2094 |
| 似然比卡方 | 3 | 4.5390 | 0.2088 |
| Mantel-Haenszel 卡方 | 1 | 0.5662 | 0.4518 |

以上给出了若干 χ^2 检验统计量，通常查看一般 χ^2 检验结果即可。

Fisher 精确检验

| | |
|---------|-----------|
| 表概率(P) | 2.971E-05 |
| Pr <= P | 0.2105 |

以上是 Fisher 精确检验的结果，当不满足一般 χ^2 检验前提条件时(SAS 输出结果会提示)，需要查看 Fisher 精确检验结果。

根据一般 χ^2 检验结果， $\chi^2 = 4.5326$ ， $P = 0.2094 > 0.05$ ，按照 0.05 的检验水准，不拒绝原假设，根据现有资料，尚不能认为汉族和回族学生中 ABO 血型分布有明显差别。

13.9 单因素多水平设计无序原因变量 $R \times 2$ 表资料统计分析

13.9.1 问题与数据

【例 13-9】 某研究为了比较三所大学大一新生中党员比例，从三所大学中随机抽取部分学生进行比较，得到表 13-8 数据。试分析三所大学新生中党员的比例差别有无统计学意义。

表 13-8 三所大学大一新生的政治面貌调查结果

| 大学 | 人 数 | | | |
|----|-------|-----|-----|-----|
| | 政治面貌： | 党员 | 非党员 | 合计 |
| A | | 53 | 73 | 126 |
| B | | 38 | 132 | 170 |
| C | | 16 | 82 | 98 |
| 合计 | | 107 | 287 | 394 |

13.9.2 对数据结构的分析

本例属于原因变量为多值名义变量、结果变量为二值变量的单因素多水平设计定性资料。

13.9.3 分析目的与统计分析方法的选择

比较原因变量各水平的频数分布情况，可以用一般 χ^2 检验或 Fisher 精确检验来处理。

13.9.4 SAS 程序中重要内容的说明

分析此资料所需的 SAS 程序如下，程序名为 SASTJFX13_9.SAS。

```
DATA SASTJFX13_9;
  DO A=1 TO 3;
    DO B=1 TO 2;
      INPUT F @@ ;
      OUTPUT;
    END;
  END;
CARDS;
53 73
38 132
16 82
;
RUN;
```

```
PROC FREQ;
  WEIGHT F;
  TABLES A* B/CHISQ FISHER;
RUN;
```

13.9.5 主要分析结果及解释

| 统计量 | 自由度 | 值 | 概率 |
|--------------------|-----|---------|--------|
| 卡方 | 2 | 21.9473 | <.0001 |
| 似然比卡方 | 2 | 21.4673 | <.0001 |
| Mantel-Haenszel 卡方 | 1 | 19.6257 | <.0001 |

以上给出了若干 χ^2 检验统计量，通常查看一般 χ^2 检验结果即可。

Fisher 精确检验

| | |
|---------|-----------|
| 表概率 (P) | 2.756E-07 |
| Pr <= P | 2.532E-05 |

以上是 Fisher 精确检验结果，当不满足一般 χ^2 检验前提条件时 (SAS 输出结果会提示)，要看 Fisher 精确检验结果。

本例资料可采用一般 χ^2 检验，得 $\chi^2 = 21.9473$ ， $P < 0.0001$ ，按照 0.05 的检验水准，拒绝原假设，接受备择假设，认为三所大学新生中的党员比例不完全相同。

13.10 单因素多水平设计有序原因变量 $R \times 2$ 表资料统计分析

13.10.1 问题与数据

【例 13-10】 为了研究吸烟年限和龋齿的关系，某人收集了如表 13-9 所示的资料。试对该资料进行统计分析。

表 13-9 吸烟年限和龋齿的关系研究结果

| 吸烟 年限 | 例 数 | | |
|----------|-----|----|----|
| | 龋齿： | 患 | 未患 |
| ≤1 | | 3 | 24 |
| 1~5 | | 9 | 28 |
| 5~10 | | 12 | 23 |
| ≥10 | | 7 | 11 |
| 合计 | | 31 | 86 |

13.10.2 对数据结构的分析

本例属于原因变量为多值有序变量、结果变量为二值变量的单因素多水平设计定性资料。

13.10.3 分析目的与统计分析方法的选择

这种情况要根据不同研究目的选择相应的统计分析方法：如果要比较不同吸烟年限的龋齿患病率是否相同，可以用一般 χ^2 检验或 Fisher 精确检验；如果要检验龋齿患病率和吸烟年限两变量之间是否呈线性趋势，可以用线性趋势检验，即 Cochran-Armitage 趋势检验。

13.10.4 SAS 程序中重要内容的说明

本例如果要作线性趋势检验，SAS 程序如下，程序名为 SASTJFX13_10.SAS。

```
DATA SASTJFX13_10;
  DO A=1 TO 4;
    DO B=1 TO 2;
      INPUT F @@ ;
      OUTPUT;
    END;
  END;
CARDS;
3 24
9 28
12 23
7 11
;
RUN;
```

```
PROC FREQ;
  WEIGHT F;
  TABLES A* B/TREND;
RUN;
```

程序中选项 TREND 要求系统进行 Cochran-Armitage 趋势检验。

13.10.5 主要分析结果及解释

Cochran-Armitage 趋势检验

| | |
|------------|--------|
| 统计量(Z) | 2.3714 |
| 单侧 Pr > Z | 0.0089 |
| 双侧 Pr > Z | 0.0177 |

以上是进行 Cochran-Armitage 趋势检验的结果，可以看出 $Z = 2.3714$ ，双侧检验 $P = 0.0177$ 。

根据 Cochran-Armitage 趋势检验的结果，拒绝原假设，接受备择假设，认为龋齿患病率与吸烟年限之间呈线性变化趋势，具体来说，患病率随着吸烟年限加大而呈增加的趋势。

13.11 单因素多水平设计双向无序 $R \times C$ 表资料统计分析

13.11.1 问题与数据

【例 13-11】某大学对计算机专业、金融专业、传媒专业各 50 名学生进行心理学测试，并判断每个学生属于哪一种典型气质类型，所得结果如表 13-10 所示。请进行合理的统计分析。

表 13-10 不同专业学生的四种气质类型分布

| 专业 | 人 数 | | | | |
|-----|-------|-----|-----|-----|-----|
| | 气质类型: | 多血质 | 胆汁质 | 抑郁质 | 黏液质 |
| 计算机 | | 16 | 13 | 7 | 14 |
| 金融 | | 12 | 15 | 10 | 13 |
| 传媒 | | 18 | 9 | 8 | 15 |
| 合计 | | 46 | 37 | 25 | 42 |

13.11.2 对数据结构的分析

本例属于原因变量和结果变量均为多值名义变量的单因素多水平设计定性资料，从列联表分类上来看属于双向无序 $R \times C$ 表。

13.11.3 分析目的与统计分析方法的选择

比较原因变量各水平的频数分布情况，可以用一般 χ^2 检验或 Fisher 精确检验来处理。

13.11.4 SAS 程序中重要内容的说明

分析此资料所需的 SAS 程序如下，程序名为 SASTJFX13_11.SAS。

```
DATA SASTJFX13_11;
  DO A=1 TO 3;
    DO B=1 TO 4;
      INPUT F @@ ; OUTPUT;
    END;
  END;
CARDS;
16 13 7 14
12 15 10 13
18 9 8 15
;
RUN;
```

```
PROC FREQ;
  WEIGHT F;
  TABLES A* B/CHISQ FISHER;
RUN;
```

13.11.5 主要分析结果及解释

| 统计量 | 自由度 | 值 | 概率 |
|--------------------|-----|--------|--------|
| 卡方 | 6 | 3.4338 | 0.7528 |
| 似然比卡方 | 6 | 3.5162 | 0.7418 |
| Mantel-Haenszel 卡方 | 1 | 0.0070 | 0.9333 |

以上给出了若干 χ^2 检验统计量，通常查看一般 χ^2 检验结果即可。

Fisher 精确检验

| | |
|---------|-----------|
| 表概率(P) | 8.189E-06 |
| Pr <= P | 0.7518 |

以上是 Fisher 精确检验结果，当不满足一般 χ^2 检验前提条件时，要看 Fisher 精确检验结果。

本例资料可采用一般 χ^2 检验, 得 $\chi^2 = 3.4338$, $P = 0.7528 > 0.05$, 按照 0.05 的检验水准, 不拒绝原假设, 根据现有资料不能认为三个专业学生的气质类型构成比例不同。

13.12 单因素多水平设计有序结果变量 $R \times C$ 表资料统计分析

13.12.1 问题与数据

【例 13-12】某研究对某市职业培训学校毕业生收入情况进行调查, 得到表 13-11 所示数据。试比较不同专业职校毕业生的收入差别有无统计学意义。

表 13-11 不同专业职校毕业生的收入情况

| 专业 | 人 数 | | | | | 合计 |
|----|--------|-------|-------------|-------------|-------|-----|
| | 月薪(元): | ≤1000 | 1000 ~ 2000 | 2000 ~ 3000 | ≥3000 | |
| 厨艺 | | 6 | 34 | 31 | 13 | 84 |
| 汽修 | | 15 | 24 | 23 | 7 | 69 |
| 财会 | | 13 | 27 | 9 | 5 | 54 |
| 合计 | | 34 | 85 | 63 | 25 | 207 |

13.12.2 对数据结构的分析

本例属于原因变量是多值名义变量、结果变量为多值有序变量的单因素多水平设计定性资料, 从列联表分类上来看属于结果变量为有序变量的单向有序 $R \times C$ 表。

13.12.3 分析目的与统计分析方法的选择

结果变量为多值有序变量, 可采用秩和检验方法来处理。

13.12.4 SAS 程序中重要内容的说明

分析此资料所需的 SAS 程序如下, 程序名为 SASTJFX13_12.SAS。

| | |
|--|--|
| DATA SASTJFX13_12; DO A=1 TO 3; DO B=1 TO 4; INPUT F @@; OUTPUT; END; END; CARDS; 6 34 31 13 15 24 23 7 13 27 9 5 ; RUN; | PROC NPAR1WAY WILCOXON; CLASS A; VAR B; FREQ F; RUN; |
|--|--|

程序与例 13-7 类似, 由于本例是多组比较, 程序执行 Kruskal-Wallis 检验。

13.12.5 主要分析结果及解释

Kruskal-Wallis Test

| | |
|-----------------|---------|
| Chi-Square | 11.0690 |
| DF | 2 |
| Pr > Chi-Square | 0.0039 |

以上是对三组进行 Kruskal-Wallis 秩和检验的结果： $\chi^2 = 11.0690, P = 0.0039 < 0.05$ 。

根据对三组进行 Kruskal-Wallis 秩和检验的结果， $P < 0.05$ ，拒绝原假设，接受备择假设，可以认为三个不同专业的职校毕业生收入有差别。相对来说，厨艺专业的毕业生收入较高，财会专业的毕业生收入较低，要想得出比较肯定的判断结果，应对三个平均秩进行两两比较，参见本书有关章节。

13.13 单因素多水平设计双向有序 $R \times C$ 表资料统计分析

13.13.1 问题与数据

【例 13-13】某矿工医院探讨矽肺不同期次患者的胸部平片密度变化，将 492 例患者资料整理成表 13-12。问：矽肺患者肺门密度的增加与期次有无关系？

表 13-12 矽肺期次与肺门密度级别的关系

| 矽肺 期次 | 例 数 | | | | 合计 |
|----------|------------|-----|-----|--|-----|
| | 肺门密度： I | II | III | | |
| I | 43 | 188 | 14 | | 245 |
| II | 1 | 96 | 72 | | 169 |
| III | 6 | 17 | 55 | | 78 |
| 合计 | 50 | 301 | 141 | | 492 |

13.13.2 对数据结构的分析

本例属于原因变量和结果变量均是多值有序变量，但属性不同的单因素多水平设计定性资料，因此从列联表分类上来看属于双向有序且属性不同的 $R \times C$ 表。

13.13.3 分析目的与统计分析方法的选择

如果要比较不同期次矽肺的肺门密度级别之间的差别是否有统计学意义，可以采用秩和检验；如果要检验矽肺期次与肺门密度两变量之间是否呈线性趋势，可以用线性趋势检验；如果要研究矽肺期次与肺门密度两变量之间是否有相关性，可以采用 Spearman 秩相关分析来处理。

13.13.4 SAS 程序中重要内容的说明

此处我们采用 Spearman 秩相关分析来分析此资料，所需的 SAS 程序如下，程序名为 SAS-TJFX13_13.SAS。

```
DATA SASTJFX13_13;
  DO A=1 TO 3;
    DO B=1 TO 3;
      INPUT F @@ ;
      OUTPUT;
    END;
  END;
CARDS;
43 188 14
1 96 72
6 17 55
;
RUN;
```

```
PROC CORR SPEARMAN;
  VAR A B;
  FREQ F;
RUN;
```

程序中调用相关过程 CORR，用 SPEARMAN 选项指定进行 Spearman 秩相关分析。

13.13.5 主要分析结果及解释

Spearman 相关系数, $N = 492$ 当 $H_0: \rho = 0$ 时, $\text{Prob} > |\mathbf{r}|$

| | A | B |
|---|-------------------|-------------------|
| A | 1.00000 | 0.53215 <.0001 |
| B | 0.53215 <.0001 | 1.00000 |

以上是 Spearman 秩相关分析的结果: $r_s = 0.53215$, $P < 0.0001$, 说明两个有序变量之间的总体相关系数不等于 0。

根据 Spearman 秩相关分析的结果, 拒绝原假设, 认为两个有序变量之间的总体相关系数不等于 0, 矽肺期次与肺门密度之间存在相关关系。结合资料可以看出, 肺门密度有随矽肺期次增高而增加的趋势。

13.14 数据库形式表达资料的统计分析

需要进行统计分析的资料通常是由一些变量和它们的观测值所组成, 计算机常用的数据结构往往是数据库形式表达的资料。这时, 数据矩阵中列上表示变量(定性或定量)的名称和具体观测值, 行上表示一个被观测对象在各变量上的具体表现(即观测值)。这种数据库结构便于进行数据录入, 但是不太容易选择合适的统计分析方法, 因为根据不同的分析目的就会利用到资料中不同的行和列信息, 从而需要选用不同的统计分析方法。尤其是判定设计类型时显得不够直观, 更不能直接看出定性变量不同水平组合下有多少频数。

【例 13-14】胃癌手术后预后因素分析数据库资料见表 13-13 和表 13-14, 结合不同分析目的选择合适的统计分析方法。

表 13-13 变量含义说明

| 变量 | 含义 | 说明 | 变量 | 含义 | 说明 |
|----|-------|----------------------------|-----|---------|-------------------------|
| x1 | 例号 | ID 号 | x8 | 手术方式 | 1 I 式 2 II 式 3 近胃 4 全切除 |
| x2 | 胃癌位置 | 1 胃底 2 胃体 3 胃窦 | x9 | 血色素 | g/L |
| x3 | 胃癌大小 | 0 ~ 5 级 | x10 | 白细胞 | 个/立方毫米 |
| x4 | 大体类型 | 1 溃疡 2 肿块 3 浸润 | x11 | 手术时年龄 | 岁 |
| x5 | 组织学类型 | 1 腺癌 2 粘液癌 3 未分化癌 4 混合型 | x12 | 性别 | 1 男性 0 女性 |
| x6 | 深度 | 1 ~ 6 级 | x13 | 是否化疗 | 1 化疗 0 未化疗 |
| x7 | 淋巴结转移 | 0 ~ 3 级 | x14 | 手术后三年情况 | 1 死亡 2 存活 |

表 13-14 胃癌手术后预后因素分析

| x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 | x9 | x10 | x11 | x12 | x13 | x14 |
|-----|-----|----|-----|----|-----|----|-----|------|------|-----|-----|-----|-----|
| 1 | 3 | 2 | 1 | 1 | 2 | 0 | 1 | 13.1 | 4408 | 35 | 1 | 1 | 2 |
| 2 | 3 | 1 | 1 | 1 | 2 | 0 | 2 | 13.4 | 5726 | 25 | 1 | 1 | 2 |
| 3 | 3 | 2 | 1 | 2 | 2 | 0 | 1 | 13.7 | 5979 | 65 | 1 | 1 | 2 |
| 4 | 3 | 3 | 1 | 1 | 6 | 3 | 2 | 8.5 | 6026 | 25 | 0 | 1 | 1 |
| ... | ... | | ... | | ... | | ... | | ... | | ... | | ... |
| 97 | 3 | 1 | 3 | 1 | 6 | 1 | 2 | 8 | 4141 | 55 | 0 | 0 | 1 |
| 98 | 3 | 2 | 1 | 1 | 6 | 0 | 2 | 10.4 | 9100 | 55 | 1 | 0 | 1 |

该资料有定量和定性两种变量，定性变量中有二值变量、多值名义变量和多值有序变量这三种情况。根据研究的需要，从定性变量中选择 1 个自变量(原因变量)和 1 个因变量(结果变量)，这时就可以组成一个标准的单因素设计二维列联表。处理这种资料的 SAS 程序与上面介绍的稍有不同，设 SAS 程序名为 SASTJFX13_14A.SAS。

| 程序 | 说明 |
|--|---|
| <pre>DATA sastjfx13_14; infile 'D:\SASTJFX\胃癌危险因素.sas'; input x1 -x14; run; ods html; proc freq; table x12* x14/ chisq fisher; run; proc npar1way wilcoxon; class x4; var x7; run; proc freq; table x7* x13/trend; run; proc corr spearman; var x3 x7; run; ods html close;</pre> | <pre>/* 数据步建立数据集 */ /* 进行卡方检验和 FISHER 精确检验 */ /* 进行秩和检验 */ /* x4 为原因变量, x7 为结果变量 */ /* 进行线性趋势检验 */ /* 进行 SPEARMAN 秩相关分析 */</pre> |

【程序说明】 因为数据很多，先将全部数据以文本格式存储在 D 盘文件夹 SASTJFX 中，文件名为：胃癌危险因素.sas。第二个 SAS 程序中包含了原始数据(文件名：SASTJFX13_14B.SAS)，占篇幅大，此处从略。

第一个过程步可对由 x12 与 x14 两个定性变量形成的二维列联表资料进行一般卡方检验和 Fisher 精确检验；第二个过程步(npar1way)可对定性变量 x4 全部水平下 x7(多值有序变量)的平均秩进行差异性检验；第三个过程步可对由 x7(多值有序变量)与 x13(二值结果变量)形成的 R×2 表资料进行线性趋势检验；第四个过程步(corr)可对由 x3 与 x7 两个有序变量形成的双向有序且属性不同的二维列联表资料进行 Spearman 秩相关分析。输出结果从略。

13.15 成组设计一元定性资料三种特殊的比较——优效性、非劣效性和等效性 t 检验

13.15.1 成组设计一元定性资料三种特殊的比较概述

成组设计定性资料率或比的三种特殊检验分别为优效性检验、非劣效性检验和等效性检验。优效性检验指主要研究目的是显示试验药的治疗效果优于对照药(安慰剂对照或阳性对照)的试验。在试验设计阶段需要设定一个界值 δ_U ，来界定试验药的优效性。非劣效性检验指主要研究目的是显示试验药的治疗效果在临床上不比阳性对照药差的试验。在试验设计阶段需要设计一个界值 δ_L ，来界定试验药是否不比阳性对照药疗效差过预先设定的这个界值。等效性检验指主要研究目的是显示两种治疗效果之间的差别大小在临床上并无重要意义的试验。在试验设计阶段需要设定等效性界值(δ_L, δ_U)来界定两种治疗的等效性。

13.15.2 优效性检验

1. 问题与数据

【例 13-15】为观察复方薤白胶囊治疗慢性支气管炎急性发作期痰热郁肺证的疗效。随机选取 240 例符合要求的病人，均分为 2 组，试验组口服复方薤白胶囊，对照组口服芩暴红止咳胶囊，疗程均为 10d。两组的有效率分别为 96.67% 和 92.50%，如果试验组的有效率超过对照组 5%，才能认为复方薤白胶囊优效于芩暴红止咳胶囊，资料见表 13-15。试评价复方薤白胶囊是否优效于芩暴红止咳胶囊。

表 13-15 两组患者的治疗效果

| 药物种类 | n | 有效率(%) |
|---------|-----|--------|
| 复方薤白胶囊 | 120 | 96.67 |
| 芩暴红止咳胶囊 | 120 | 92.50 |

2. 如何用 SAS 实现计算

【分析与解答】该资料的研究目的是评价复方薤白胶囊是否优效于芩暴红止咳胶囊，并且设定了优效性的界值，所以应该采用的方法是优效性检验。

下面用 SAS 软件编程法进行分析，程序名为 SASTJFX13_15.SAS。

```
data SASTJFX13_15;
    n1=120;n2=120;          /* 1 */
    p1=0.9667;p2=0.9250;    /* 2 */
    U=0.05;                  /* 3 */
    v_p1=p1*(1-p1)/n1;      /* 4 */
    v_p2=p2*(1-p2)/n2;      /* 5 */
    se=sqrt(v_p1+v_p2);     /* 6 */
    t=((p1-p2)-U)/se;       /* 7 */
    p=1-probt(t,n1+n2-2);   /* 8 */
    ods html;
    proc print;              /* 9 */
        var t p;
    run;
    ods html close;
```

【程序说明】 data 步建立数据集 DA14_9，第 1 步定义试验组和对照组的样本量；第 2 步定义试验组和对照组的有效率；第 3 步定义优效性检验的界值；第 4 步和第 5 步分别计算两个样本率的方差；第 6 步计算两个样本率之差的标准差；第 7 步计算检验统计量 t 值；第 8 步计算 t 值对应的 t 分布右侧的累积概率；第 9 步输出 t 和 p 值。

【主要输出结果及解释】 输出结果为：

| Obs | t | p |
|-----|----------|---------|
| 1 | -0.28530 | 0.61217 |

【统计和专业结论】 $t = -0.28530$, $p = 0.61217$, 按照 $\alpha = 0.05$, 不能拒绝 H_0 , 还不能认为复方薤白胶囊在治疗慢性支气管炎急性发作期痰热郁肺证的有效性优于芩暴红止咳胶囊。

13.15.3 非劣效性检验

1. 问题与数据

【例 13-16】为评价度洛西汀肠溶胶囊治疗抑郁症的疗效，采用随机、双盲、氟西汀平行对照的方法进行试验。受试者分别口服度洛西汀肠溶胶囊或氟西汀胶囊，共观察 6 周。其中，试验组 111 例患者，对照组 117 例患者。治疗 6 周观察疗效，试验组和对照组的有效率分别为 88.0% 和 89.2%，资料见表 13-16。根据临床实际，设置非劣效性界限为 -10%。试评价度洛西汀肠溶胶囊治疗抑郁症的疗效是否非劣效于氟西汀。

表 13-16 两组患者的治疗效果

| 药物种类 | n | 有效率(%) |
|----------|-----|--------|
| 度洛西汀肠溶胶囊 | 111 | 88.0 |
| 氟西汀胶囊 | 117 | 89.2 |

2. 如何用 SAS 实现计算

【分析与解答】 该资料属于成组设计定性资料，研究目的是评价度洛西汀肠溶胶囊治疗抑郁症的疗效是否非劣效于氟西汀，设定非劣效性界值为 -0.10 ，使用非劣效性检验分析该资料。

下面用 SAS 软件编程法进行分析，程序名为 SASTJFX13_16. SAS。

```
data SASTJFX13_16;
  n1 = 111; n2 = 117;
  p1 = 0.880; p2 = 0.892;
  L = -0.10; /* 1 */
  v_p1 = p1 * (1 - p1) / n1;
  v_p2 = p2 * (1 - p2) / n2;
  se = sqrt(v_p1 + v_p2);

  t = (p1 - p2) - L / se; /* 2 */
  p = 1 - probt(t, n1 + n2 - 2);
ods html;
proc print;
  var t p;
run;
ods html close;
```

【程序说明】 第 1 步定义非劣效性检验的界值；第 2 步对 H_0 ，即非劣效性界值 L 进行假设检验，计算 t 值。

【主要输出结果及解释】 输出结果为：

| Obs | t | p |
|-----|---------|----------|
| 1 | 2.08889 | 0.018918 |

【统计和专业结论】 $t = 2.08889$ ， $p = 0.018918$ ，按照 $\alpha = 0.05$ ，拒绝 H_0 ，接受 H_1 ，可以认为度洛西汀肠溶胶囊在治疗抑郁症方面非劣效于氟西汀。

13.15.4 等效性检验

1. 问题与数据

【例 13-17】 一个新的抗肿瘤药物 A 与临床有效药物 B 对照进行临床试验，选取 300 例符合要求的病人随机均分成两组，分别接受 A 药和 B 药治疗。试验结果为 A 药的有效率是 58.0%，B 药的有效率是 46.0%，资料见表 13-17。根据临床实际，设置等效性界限为 10%，问：两种药物是否等效？

2. 如何用 SAS 实现计算

【分析与解答】 该资料属于成组设计定性资料，分析目的是评价两种药物是否等效，应该采用双单侧检验进行等效性检验，设定等效性界值为 $L = -0.10$ ， $U = 0.10$ 。

表 13-17 两组患者的治疗效果

| 药物种类 | n | 有效率 |
|------|-----|-------|
| A | 150 | 58.0% |
| B | 150 | 46.0% |

下面用 SAS 软件编程法进行分析，程序名为 SASTJFX13_17. SAS。

```
data SASTJFX13_17;
  n1 = 150; n2 = 150;
  p1 = 0.58; p2 = 0.46;
  L = -0.10; U = 0.10;
  v_p1 = p1 * (1 - p1) / n1;
  v_p2 = p2 * (1 - p2) / n2;
  se = sqrt(v_p1 + v_p2);
  t1 = (p1 - p2) - L / se; /* 1 */
  t2 = (U - (p1 - p2)) / se; /* 2 */

  p1 = 1 - probt(t1, n1 + n2 - 2); /* 3 */
  p2 = 1 - probt(t2, n1 + n2 - 2); /* 4 */
ods html;
proc print;
  var t1 p1 t2 p2;
run;
ods html close;
```

【程序说明】 第 1 步对 $H_{0(1)}$, 即下限 L 进行假设检验, 计算 t_1 ; 第 2 步对 $H_{0(2)}$, 即上限 U 进行假设检验, 计算 t_2 ; 第 3 步计算 t_1 对应的 t 分布右侧的累积概率; 第 4 步计算 t_2 对应的 t 分布右侧的累积概率。

【主要输出结果及解释】 输出结果为:

| Obs | t1 | p1 | t2 | p2 |
|-----|---------|------------|----------|---------|
| 1 | 3.84137 | .000074756 | -0.34922 | 0.63641 |

$t_1 = 3.84137$, $p_1 = 0.000074756$, 按照 $\alpha = 0.025$, 拒绝 $H_{0(1)}$, 接受 $H_{1(1)}$; $t_2 = -0.34922$, $p_2 = 0.63641$, 按照 $\alpha = 0.025$, 不能拒绝 $H_{0(2)}$ 。只有两个单侧检验均拒绝 H_0 , 才可以认为两种药物等效。

【统计和专业结论】 因为第一个单侧检验没有拒绝 H_0 , 所以还不能认为两种抗肿瘤药物是等效的。

13.16 本章小结

本章对单因素设计定性资料(即通常所说的各种二维列联表资料)的统计分析方法和 SAS 实现技术进行了全面介绍, 还对以数据库形式呈现的、结果变量为定性变量的资料如何用 SAS 实现统计分析进行了扼要介绍。

(毛玮 武佳 张泽 陶丽新)

第 14 章 多因素设计一元定性资料差异性分析

本章主要介绍利用 SAS 软件分析多因素设计一元定性资料的方法，包括加权 χ^2 检验、CMH 检验、meta 分析和 ROC 分析。为节省篇幅，书中仅给出程序及其说明、主要输出结果及结果解释。

14.1 用加权 χ^2 检验处理结果变量为二值变量的高维列联表资料

14.1.1 问题与数据

【例 14-1】 欲考察两种药物（盆腔炎、野菊花栓）治疗慢性盆腔炎的疗效，某研究者经过临床试验后收集到的资料如表 14-1 所示，请选用较简便的统计分析方法评价两种药物的效果并尽量消除“病情轻重”对结果的影响。

表 14-1 盆腔炎和野菊花栓治疗慢性盆腔炎疗效的观察结果

| 病情程度 | 药物种类 | 例 数 | |
|------|------|-----|----|
| | | 有效 | 无效 |
| 轻 | 盆腔炎 | 87 | 2 |
| | 野菊花栓 | 28 | 6 |
| 中 | 盆腔炎 | 48 | 6 |
| | 野菊花栓 | 136 | 10 |
| 重 | 盆腔炎 | 72 | 5 |
| | 野菊花栓 | 13 | 7 |

14.1.2 对数据结构的分析

由表 14-1 可以看出，该列联表资料涉及两个原因变量，即“药物种类”、“病情程度”；一个结果变量，即“疗效”，它是一个二值结果变量，故该资料是结果变量为二值变量的三维列联表资料。

14.1.3 分析目的与统计分析方法的选择

该研究分析目的是评价两种药物的效果并尽量消除“病情轻重”对结果的影响。在三维列联表中，通常有两个原因变量和一个结果变量，不同的研究目的决定了选用不同的统计分析方法，对于结果变量为二值变量的高维列联表，可选用加权 χ^2 检验、CMH χ^2 检验、多重 logistic 回归、对数线性模型等。若不想用复杂的对数线性模型或 logistic 回归模型来分析三维列联表资料，并且资料又不适合采用简单“合并”方式处理时，就可采用加权 χ^2 检验（消除掉一个原因变量对结果变量的影响，考察另一个原因变量与结果变量之间是否独立）、CMH χ^2 检验（消除掉一个原因变量对结果变量的影响，计算优势比 OR 或相对危险度 RR 并对其进行假设检验）。这两种检验方法都无法回答被合并掉的那个原因变量对结果变量的影响作用有多大，只是对其进行分层计算，即评价另一个原因变量对结果变量的影响时将其对结果变量的影响扣除掉。本例介绍如何基于计算公式实现加权 χ^2 检验和 MH χ^2 检验（即 CMH χ^2 检验），下一节将介绍如何用 SAS 中 FREQ 过程直接实现 CMH χ^2 检验，其他方法请读者参阅本书相关章节。

14.1.4 SAS 程序中重要内容的说明

设分析例 14-1 定性资料的程序名为 SASTJFX14_1.SAS。

```
DATA SASTJFX14_1;
DO k=1 TO 3;
INPUT a b c d;
n=a+b+c+d; e=a+b; f=c+d; g=a
+c; h=b+d;
d1=(a*d-b*c)/n; i=c+d; d2=e*
f/n; pd=e*f*g*h; v1=pd/n**3;
k2=pd/(n-1)/n**2; j=a*f; k=c
*e;
l=j/n; m=k/n; OUTPUT; END; CARDS;
87 2 28 6
48 6 136 10
72 5 13 7
;
RUN;
ods html;
PROC MEANS NOPRINT;
VAR d1 d2 v1 k2 l m;
OUTPUT OUT=aaa SUM=d1 d2 v1 k2 l m;
RUN;
```

```
DATA a2; set aaa;
v2=d2**2; md=d1/d2; v=v1/v2; k1
=d1**2;
wchisq=ROUND(md**2/v,0.001); wp
=1-PROBCHI(wchisq,1);
rr=round(l/m,0.001); mhchisq=
ROUND(k1/k2,0.001);
mhp=1-PROBCHI(mhchisq,1);
file print;
put #3 @5 'W-chisq=' wchisq @35 'W-
p=' wp
#6 @5 'RR=' rr @15 'MH-chisq='
mhchisq @35 'MHW-p=' mhp;
RUN;
ods html close;
```

数据步中的公式是计算加权 χ^2 、MH χ^2 统计量及 RR 值的过程。

14.1.5 主要分析结果及解释

$$\begin{aligned} W\text{-chisq} &= 7.194 & W\text{-p} &= 0.0073147742 \\ RR &= 1.099 & MH\text{-chisq} &= 7.142 & MH\text{-p} &= 0.0075299133 \end{aligned}$$

W-chisq 为加权 χ^2 值, W-p 为对应的 P 值, 即加权 $\chi^2 = 7.194$, $P = 0.0073$; 相对危险度为 $RR = 1.099$, MH-chisq 为 MH χ^2 , MHW-p 为其对应的 P 值, 即 $MH \chi^2 = 7.142$, $P = 0.0075$ 。

专业结论: 消除病情的影响后, 发现两种药物的有效率之间的差别有统计学意义。盆炎栓组的有效率为 $P_e(a/a+b) \approx (207/220) \times 100\% = 94.09\%$, 野菊花栓组有效率为 $P_0(c/c+d) \approx (177/200) \times 100\% = 88.50\%$ (注: 这两个有效率是简单合并计算而得, 只是一个近似值), 由加权 χ^2 检验结果可知, 两个有效率之间的差别有统计学意义, 盆炎栓组的有效率高于野菊花栓组的有效率; 由 CMH χ^2 检验结果可知, $RR = P_e/P_0 = 1.099$, 即盆炎栓组的有效率是野菊花栓组的有效率的 1.099 倍, 此值与 1 之间的差别有统计学意义, 而是否有临床意义应根据临床专家的经验而定, 从绝对数值上来看, 似乎实际意义并不明显。

14.2 用 CMH χ^2 检验处理结果变量具有三种性质的高维列联表资料

14.2.1 问题与数据

【例 14-2】 为了考察吸烟与死亡率之间的关系, 对一些吸烟者和不吸烟者进行为期 20 年的随访观察, 最终记录他们的生存状态, 按年龄分层后, 各年龄组的生存情况见表 14-2, 试对该资料进行分析。

表 14-2 不同年龄吸烟者与不吸烟者生存情况

| 年龄(岁) | 吸烟状态 | 例 数 | | 年龄(岁) | 吸烟状态 | 例 数 | |
|---------|------|-------|--------|---------|------|-------|--------|
| | | 生存情况: | 生存 死亡 | | | 生存情况: | 生存 死亡 |
| 18 ~ 24 | 吸烟者 | | 53 2 | 55 ~ 64 | 吸烟者 | | 64 51 |
| | 不吸烟者 | | 61 1 | | 不吸烟者 | | 81 40 |
| 25 ~ 34 | 吸烟者 | | 121 3 | 65 ~ 74 | 吸烟者 | | 7 29 |
| | 不吸烟者 | | 152 5 | | 不吸烟者 | | 28 101 |
| 35 ~ 44 | 吸烟者 | | 95 14 | ≥75 | 吸烟者 | | 0 13 |
| | 不吸烟者 | | 114 7 | | 不吸烟者 | | 0 64 |
| 45 ~ 54 | 吸烟者 | | 103 27 | | | | |
| | 不吸烟者 | | 66 12 | | | | |

【例 14-3】 在一项治疗小细胞肺癌的临床试验中，患者被随机的分为两组，连续治疗组在每一个治疗周期使用相同的化学药物联合疗法，交替治疗组在不同治疗周期使用不同的药物组合，不同性别患者的治疗结果见表 14-3。试对两种疗法的疗效做出评价。

表 14-3 两种疗法治疗小细胞肺癌患者的疗效情况

| 性别 | 治疗方法 | 例 数 | | | |
|----|------|-----|------|-----|-----------|
| | | 疗效: | 疾病发展 | 无变化 | 部分缓解 完全缓解 |
| 男性 | 连续治疗 | | 28 | 45 | 29 26 |
| | 交替治疗 | | 41 | 44 | 20 20 |
| 女性 | 连续治疗 | | 4 | 12 | 5 2 |
| | 交替治疗 | | 12 | 7 | 3 1 |

【例 14-4】 研究不同社区、性别的成人获取健康知识途径的差别，调查数据见表 14-4。

表 14-4 不同社区、性别的成人获取健康知识途径的差别

| 社区 | 性别 | 例 数 | | | |
|----|----|-----|------|----|------|
| | | 途径: | 大众媒体 | 网络 | 社区教育 |
| 1 | 男 | | 20 | 35 | 26 |
| | 女 | | 10 | 27 | 57 |
| 2 | 男 | | 42 | 17 | 26 |
| | 女 | | 16 | 12 | 26 |
| 3 | 男 | | 15 | 15 | 16 |
| | 女 | | 11 | 12 | 20 |

14.2.2 对数据结构的分析

由表 14-2 可以看出，该列联表资料涉及了两个原因变量：“年龄”和“吸烟状态”，一个结果变量：“生存情况”。也就是说，该资料属于结果变量为二值变量的三维列联表资料。

由表 14-3 可以看出，原因变量包括“性别”和“治疗方法”，本资料结果变量为“疗效”，是一个有序变量，所以，本资料是结果变量为多值有序变量的三维列联表资料。

由表 14-4 可以看出，原因变量为“社区”和“性别”，其中“社区”为多值名义变量；“性别”为二值名义变量；结果变量为获取知识的途径，也属于多值名义变量，有三个水平，分别为大众媒体、网络和社区教育。所以该资料属于结果变量为多值无序(或名义)变量的三维列联表资料。

14.2.3 分析目的与统计分析方法的选择

例 14-2 的研究分析目的是研究吸烟者与不吸烟者生存概率之间的差别是否具有统计学意义。由于这是结果变量为二值变量的三维列联表，故可以选用加权 χ^2 检验(见例 14-1 及其分析和结果解释)、CMH χ^2 检验、logistic 回归和对数线性模型。CMH 统计分析(Cochran-Mantel-

Haenszel statistics)是在 MH 统计分析方法的基础上发展并提出来的,现在统称为扩展的 MH 卡方统计量,也统称为 MH 检验,用于分层分析即控制混杂因素后对二维列联表资料的统计处理。

例 14-3 的研究分析目的是研究两种治疗方法所得疗效之间的差别是否具有统计学意义,但应当消除性别对结果的影响。由于本资料属于结果变量为多值有序变量的高维列联表资料,可以选用 CMH χ^2 检验(即 CMH 校正的秩和检验)和有序变量 logistic 回归分析进行处理。对数线性模型无法利用资料的有序性,因此不宜选用。若采用有序变量多重 logistic 回归分析,应注意结合原因变量中是否存在多值名义变量或多值有序变量而决定对原因变量的赋值方法。

例 14-4 的研究分析目的是考察不同社区、性别对获取健康知识途径的差别,由于本资料属于结果变量为多值无序变量的高维列联表资料,故可以采用 CMH χ^2 检验、扩展的 logistic 回归分析和对数线性模型。本例中原因变量为社区和性别,其中社区为多值名义变量;性别为二值名义变量;结果变量为获取知识的途径,也属于多值名义变量。现在想要考察不同社区、性别对获取健康知识途径的差别,可以采用 CMH χ^2 检验。

值得注意的是,CMH χ^2 检验包含三种检验方法:(1)非零相关检验(适合于原因变量与结果变量都是多值有序变量);(2)行平均得分差检验(仅考察原因变量全部水平组之间在结果上的差别是否具有统计学意义,结果变量必须是多值有序变量);(3)一般关联性检验(适合于原因变量和结果变量都是名义变量)。这里所提及的“原因变量”是指在多个原因变量中被保留下来的那个原因变量。

14.2.4 SAS 程序中重要内容的说明

设分析例 14-2 定性资料的程序名为 SASTJFX14_2.SAS。

| | |
|--|--|
| <pre>data SASTJFX14_2; do age =1 to 7; do smoking =1 to 2; do outcome =1 to 2; input f@@ ;output; end;end;end; cards; 53 2 61 1 121 3 152 5 95 14 114 7 103 27 66 12 64 51 81 40 7 29 28 101 0 13 0 64 ; run;</pre> | <pre>proc freq data =SASTJFX14_2; tables age* smoking* outcome/cmh; weight f; ods html; run;</pre> |
|--|--|

首先建立数据集,程序中的 age、smoking 和 outcome 分别代表年龄、吸烟状态和生存情况,变量 f 表示频数。数据的分析采用 FREQ 过程,SAS 语句的写法与二维表大体一致,不同之处在于 tables 语句中的变量个数,这里是三维表,需要列出三个变量,依次为年龄、吸烟状态和生存情况,列在第一位的变量是需要控制的原因变量,列在第二位的变量是想要考察的原因变量,列在第三位的变量是结果变量。需要注意的是,若在程序中将变量 age 和 smoking 的位置调换一下,得到的结果是不一样的,这时考察的目的是平衡了吸烟状态的影响之后,各年龄层对应的生存率之间的差别是否具有统计学意义。tables 语句中的选项 cmh 指定输出 CMH 统计量。ods html 语句则要求以 HTML 格式输出结果。

设分析例 14-3 定性资料的程序名为 SASTJFX14_3.SAS。

```
data SASTJFX14_3;
do gender=1 to 2;
do treat=1 to 2;
do lx=1 to 4;
input f@@ ;output;
end;end;end;
cards;
28 45 29 26 41 44 20 20
4 12 5 2 12 7 3 1
;
run;
```

```
proc freq data = SASTJFX14_3;
tables gender* treat* lx/cmh2;
weight f;
ods html;
run;
```

程序中的 gender、treat 和 lx 分别代表性别、治疗方法和疗效，变量 f 表示频数。FREQ 过程步的语句与例 14-2 基本相同，不同的只是在 tables 语句中使用了 cmh2 选项，该选项指定输出前两个 CMH 统计量，即相关统计量和行平均得分统计量，排除了一般关联统计量。当然，此处仍然可以使用 cmh 选项。

设分析例 14-4 定性资料的程序名为 SASTJFX14_4.SAS。

```
Data SASTJFX14_4;
do community=1 to 3;
do sex=0 to 1;
do way=1 to 3;
input f@@ ;
output;
end;
end;
end;
cards;
20 35 26
10 27 57
42 17 26
16 12 26
15 15 16
11 12 20
;
run;
```

```
proc freq data = SASTJFX14_4;
tables community* sex* way/cmh;
weight f;
ods html;
run;
```

用 community 代表社区；sex 表示性别，其取值 0 和 1 分别代表男性和女性；way 代表获取健康知识的途径，其取值 1、2、3 分别表示大众媒体、网络和社区教育。tables 后面依次写两个原因变量和结果变量，这里需要注意结果变量必须写在最后，分析时是按第一个原因变量，也就是社区进行分层。此外，也可以要求输出其他一些统计量。

14.2.5 主要分析结果及解释

例 14-2 的输出结果：

| Cochran-Mantel-Haenszel Statistics (Based on Table Scores) | | | | |
|--|------------------------|----|--------|--------|
| Statistic | Alternative Hypothesis | DF | Value | Prob |
| 1 | Nonzero Correlation | 1 | 5.8443 | 0.0156 |
| 2 | Row Mean Scores Differ | 1 | 5.8443 | 0.0156 |
| 3 | General Association | 1 | 5.8443 | 0.0156 |

Estimates of the Common Relative Risk (Row1/Row2)

| Type of Study | Method | Value | 95% Confidence Limits | |
|---------------|-----------------|--------|-----------------------|--------|
| Cohort | Mantel-Haenszel | 1.2131 | 1.0372 | 1.4187 |
| (Col2 Risk) | Logit | 1.1336 | 0.9740 | 1.3194 |

$\chi^2_{CMH} = 5.8443$, $P = 0.0156$ 。说明消除年龄影响后,吸烟与否对应的生存率是不同的,由实际数据得知:吸烟者较不吸烟者生存率低。CMH 检验中的 $P < 0.05$, 拒绝原假设,接受备择假设,说明在控制了年龄因素的影响之后,吸烟者与不吸烟者的生存情况有统计学差异,可以认为吸烟者的死亡率要高于不吸烟者;相对危险度为 1.2131,说明吸烟者的死亡率是不吸烟者的 1.2131 倍。

例 14-3 的输出结果:

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

| Statistic | Alternative Hypothesis | DF | Value | Prob |
|-----------|------------------------|----|--------|--------|
| 1 | Nonzero Correlation | 1 | 6.3833 | 0.0115 |
| 2 | Row Mean Scores Differ | 1 | 6.3833 | 0.0115 |

资料类型属于结果变量为多值有序变量的三维列联表资料,同时原因变量是二值的,所以使用行平均得分统计量,此处自由度 $\nu = 1$, $\chi^2_{CMH} = 6.3833$, $P = 0.0115$ 。CMH 检验中的 $P < 0.05$, 拒绝原假设,说明在控制了性别因素的基础上,连续治疗组与交替治疗组的疗效有统计学差异,可以认为连续治疗组的疗效要优于交替治疗组。

例 14-4 的输出结果:

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

| Statistic | Alternative Hypothesis | DF | Value | Prob |
|-----------|------------------------|----|---------|--------|
| 1 | Nonzero Correlation | 1 | 18.6284 | <.0001 |
| 2 | Row Mean Scores Differ | 1 | 18.6284 | <.0001 |
| 3 | General Association | 2 | 19.3826 | <.0001 |

结果给出了 CMH 统计量,这些统计量都是在使用 table 方法打分的基础上计算的,本例需要看第三个统计量,即一般关联 (General Association) 统计量,此处自由度为 2, $\chi^2_{CMH} = 19.3826$, $P < 0.0001$, 拒绝原假设 H_0 , 接受 H_1 , 说明至少有一层,原因变量和结果变量之间存在某种关联。

还有一点需要注意的是,在程序中将变量 community 和 sex 的位置调换一下,得到的结果是不一样的,这时考察的目的是平衡了性别的影响之后,社区与获取健康知识的途径之间有无关联。请读者自行尝试。

专业结论: 在控制了社区这个因素之后,性别与获取健康知识的途径存在一定的关联。结合列联表中的行百分数可以看到,女性较多地通过社区教育的方式获得健康知识,而男性则更多地依赖于大众媒体和网络这两种途径。

14.3 用 meta 分析分别合并处理多个成组设计定性资料

14.3.1 问题与数据

【例 14-5】 有 7 项关于阿司匹林预防心肌梗死后死亡的随机对照实验研究,结果见表 14-5。试对该资料进行 meta 分析。

表 14-5 阿司匹林预防心肌梗死后死亡
的7项随机对照实验研究结果

| 研究编号 | 观察人数(死亡人数) | | 研究编号 | 观察人数(死亡人数) | |
|------|------------|----------|------|------------|------------|
| | 阿司匹林组 | 安慰剂组 | | 阿司匹林组 | 安慰剂组 |
| S1 | 615(49) | 624(67) | S5 | 810(85) | 406(52) |
| S2 | 758(44) | 771(64) | S6 | 2267(246) | 2257(219) |
| S3 | 832(102) | 850(126) | S7 | 8578(1570) | 8600(1720) |
| S4 | 317(32) | 309(38) | | | |

【例 14-6】 为了探讨低频电磁场(ELF)与儿童白血病的关系，收集与此有关的 9 项病例-对照研究，结果见表 14-6。请进行 meta 分析。

表 14-6 ELF 与儿童白血病关系的 9 项病例-对照研究结果

| 研究编号 | 病例组人数 | | 对照组人数 | | OR | 95%CI | P |
|------|-------|-----|-------|-----|------|--------------|--------|
| | 高暴露 | 低暴露 | 高暴露 | 低暴露 | | | |
| 1 | 63 | 92 | 29 | 126 | 2.98 | 1.72 - 5.15 | 0.0050 |
| 2 | 103 | 95 | 113 | 112 | 1.08 | 0.72 - 1.60 | 0.7210 |
| 3 | 25 | 218 | 17 | 195 | 1.32 | 0.66 - 2.63 | 0.4042 |
| 4 | 18 | 162 | 25 | 252 | 1.12 | 0.56 - 2.21 | 0.7273 |
| 5 | 27 | 70 | 52 | 207 | 1.54 | 0.87 - 2.72 | 0.1168 |
| 6 | 14 | 70 | 15 | 126 | 1.68 | 0.72 - 3.94 | 0.1918 |
| 7 | 29 | 52 | 16 | 61 | 2.13 | 0.914 - 4.62 | 0.0365 |
| 8 | 122 | 89 | 92 | 113 | 1.68 | 1.12 - 2.53 | 0.0083 |
| 9 | 49 | 55 | 30 | 74 | 2.20 | 1.19 - 4.06 | 0.0066 |

【例 14-7】 某研究者收集到关于某种诊断方法的满足纳入条件的 10 项独立的诊断试验结果(如表 14-7 所示)。若各研究所采用的参考试验为严格的金标准，试进行 meta 分析。

表 14-7 10 项诊断试验的结果

| 研究编号 (ID) | 真阳性例数 (nTP) | 假阳性例数 (nFP) | 假阴性例数 (nFN) | 真阴性例数 (nTN) | 灵敏度 (Se) | 特异度 (Sp) |
|--------------|----------------|----------------|----------------|----------------|-------------|-------------|
| 1 | 30 | 2 | 3 | 55 | 0.91 | 0.96 |
| 2 | 22 | 2 | 0 | 25 | 1.00 | 0.93 |
| 3 | 56 | 3 | 3 | 89 | 0.95 | 0.97 |
| 4 | 53 | 11 | 7 | 132 | 0.88 | 0.92 |
| 5 | 30 | 15 | 14 | 95 | 0.69 | 0.86 |
| 6 | 4 | 0 | 2 | 18 | 0.67 | 1.00 |
| 7 | 9 | 3 | 7 | 28 | 0.57 | 0.89 |
| 8 | 39 | 7 | 4 | 25 | 0.91 | 0.78 |
| 9 | 7 | 1 | 4 | 8 | 0.63 | 0.89 |
| 10 | 18 | 1 | 0 | 23 | 1.00 | 0.96 |

14.3.2 对数据结构的分析

由表 14-5 可以看出，各项研究所对应的实验设计类型均为队列研究设计。队列研究设计是依据专业知识，通过对不同暴露水平的对象进行追踪观察，确定其疾病发生情况，从而分析暴露因素与疾病发生之间的因果关系，故此表资料本质上为 7 个队列研究设计的四格表(或 2×2 表)资料。

由表 14-6 可以看出, 各项研究所对应的实验设计类型均为病例 - 对照研究设计。该设计是指在要了解暴露于某种因素对疾病的发生有无影响及其影响程度时, 针对某因素从部分病人发病之后开始调查, 将病人设为病例组, 并选择其他条件与病例接近的非病人设为对照组, 分别调查这两组人暴露于可疑致病因子的情况, 故此表资料本质上为 9 个病例 - 对照研究设计的四格表资料。

由表 14-7 可以看出, 各项研究所对应的实验设计类型均为诊断试验设计。诊断试验设计是指按照配对原则分别接受两种不同的处理方法, 每种处理方法的结果都可分为“阳性”和“阴性”两种, 故此表资料本质上为 10 个诊断试验四格表资料(或 2×2 配对设计定性资料)。

14.3.3 分析目的与统计分析方法的选择

在科研中, 针对同一问题常常同时或者先后有许多类似的研究。然而, 由于研究对象、设计方案、干预措施、结局变量、样本含量、随访时间、各种干扰因素(即非试验因素)的影响、研究过程中质量控制的严格程度等方面并不完全相同, 致使各研究结果也不完全一致, 有些研究结果甚至相互矛盾。当出现针对同一问题的各项研究结果不一致甚至相反时, 人们究竟该相信哪一项研究的结论呢? 要回答这一问题, 有两种办法: 一是通过严格设计的大规模随机对照试验进行验证; 二是通过对这些研究及其结果进行综合分析和再评价。meta 分析就是后一种办法中经常要用到的用于定量合成的一种统计学处理方法。

由于例 14-6 ~ 例 14-8 的分析目的是比较同一研究的不同结果的同质性, 所以可根据各自的实验设计类型选择单因素 2 水平设计定量资料、队列研究设计四格表资料、病例-对照研究设计四格表资料和诊断试验四格表资料 meta 分析。

14.3.4 SAS 程序中重要内容的说明

设分析例 14-5 定性资料的程序名为 SASTJFX14_5.SAS。

```

OPTIONS LS = 74 PS = MAX CENTER NO-
DATE;
DATA A;
INPUT study_ID ni1 n_ai ni2 n_ci @@ ;
n_bi = ni1 - n_ai; n_di = ni2 - n_ci;
n_ai + 0.25; n_bi + 0.25; n_ci + 0.25;
n_di + 0.25;
RR = (n_ai / (n_ai + n_bi)) / (n_ci /
(n_ci + n_di));
yi = LOG(RR);
wi = (1/n_ai - 1/(n_ai + n_bi) + 1/n_ci -
1/(n_ci + n_di)) * (-1);
Ni = n_ai + n_bi + n_ci + n_di;
number_of_study = 7;
CARDS;
1 615 49 624 67
2 758 44 771 64
3 832 102 850 126
4 317 32 309 38
5 810 85 406 52
6 2267 246 2257 219
7 8578 1570 8600 1720
w_bar = sum_of_wi / number_of_study;
Square_of_s_w = (Sum_of_wi2 - number_
_of_study * w_bar * 2) / (number_of_
study - 1);
U = (number_of_study - 1) * (w_bar -
Square_of_s_w / (number_of_study * w_
bar));
Square_of_tao_hat = (Q - (number_of_
study - 1)) / U;
IF Square_of_tao_hat < 0 THEN Square_
_of_tao_hat = 0; END;
RUN;
PROC SORT; BY DESCENDING ID;
DATA C; SET B;
Square_of_tao_hat_for_all + Square_
_of_tao_hat;
wi_bar = 1 / (1/wi + Square_of_tao_hat_
_for_all);
RUN;
PROC SORT; BY ID;
DATA D; SET C;
Sum_of_wi_bar_temp + wi_bar;

```

```

;
RUN;
DATA B; SET A;
wi2=wi** 2; wiyi=wi* yi; wiyi2 =
    wi* yi** 2;
sum_of_wi_temp+wi; sum_of_wi2_temp
+wi2;
sum_of_wiyi_temp+wiyi;
sum_of_wiyi2_temp+wiyi2;
ID=_N_;
IF _N_=number_of_study THEN DO;
sum_of_wi=sum_of_wi_temp;
sum_of_wi2=sum_of_wi2_temp;
sum_of_wiyi=sum_of_wiyi_temp;
sum_of_wiyi2=sum_of_wiyi2_temp;
yw_bar=sum_of_wiyi/sum_of_wi;
square_of_s_y_bar=1/sum_of_wi;
Q=sum_of_wiyi2 - yw_bar** 2* sum
of_wi;
DF=number_of_study -1;
P=1 - PROBCHI (Q, DF);
I2 = (Q - (number_of_study - 1))/Q
* 100;
IF I2 < 0 THEN I2 = 0;
FRRRC = EXP (yw_bar);
FRRLOW = EXP (yw_bar - 1.96* square
of_s_y_bar** 0.5);
FRRUP = EXP (yw_bar + 1.96* square_of
_s_y_bar** 0.5);

Sum_of_wi_baryi_temp+wi_bar* yi;
IF _N_=number_of_study THEN DO;
Sum_of_wi_bar=Sum_of_wi_bar_temp;
Sum_of_wi_baryi = Sum_of_wi_baryi
temp;
V_hat_y_bar_hat=1/Sum_of_wi_bar;
y_bar_hat = Sum_of_wi_baryi/Sum_of
wi_bar;
RRC = EXP (y_bar_hat);
RRLOW = EXP (y_bar_hat - 1.96* V_hat
y_bar_hat** 0.5);
RRUP = EXP (y_bar_hat + 1.96* V_hat_y
_bar_hat** 0.5); END;
RUN;
PROC SORT; BY DESCENDING ID;
DATA E; SET D; WHERE ID = number_of
study; RUN;
PROC PRINT;
VAR number_of_study Q DF P I2 FRRRC
FRRLOW FRRUP RRC RRLOW RRUP;
ODS HTML;
RUN;

```

程序中分别用 n_{i1} 、 n_{i2} 表示各研究中试验组和对照组的样本含量；用 N_i 表示各研究的样本含量；分别用 n_{ai} 、 n_{bi} 、 n_{ci} 、 n_{di} 表示各研究对应四格表 4 个格子上的频数；用 y_i 表示相对危险度 RR 的对数值；用 w_i 表示采用固定效应模型时的权重；用 $wi2$ 表示 w_i 的平方；用 $wiyi$ 表示 w_i 与 y_i 的乘积；用 $wiyi2$ 表示 w_i 与 y_i 的平方的乘积；用 $number_of_study$ 表示纳入研究的数量；用 $Sum_of_wi_temp$ 、 $Sum_of_wi2_temp$ 、 $Sum_of_widi_temp$ 和 $Sum_of_widi2_temp$ 分别表示 w_i 、 $wi2$ 、 $widi$ 和 $widi2$ 累积的中间结果，用 Sum_of_wi 、 Sum_of_wi2 、 Sum_of_widi 和 Sum_of_widi2 分别表示 w_i 、 $wi2$ 、 $widi$ 和 $widi2$ 的累积的最终结果；用 yw_bar 和 y_bar_hat 分别表示采用固定效应模型和随机效应模型时的合并效应；用 w_bar 、 $Square_of_s_w$ 、 $Square_of_tao_hat$ 和 wi_bar 分别表示 \bar{w} 、 s_w^2 、 $\hat{\tau}^2$ 和 \bar{w}_i ；用 DF 表示自由度；用 Q 和 P 分别表示异质性检验的 χ^2 值和 P 值； $I2$ 为异质性检验的另一个统计量；用 $Sum_of_wi_bar_temp$ 和 $Sum_of_wi_baryi_temp$ 分别表示 wi_bar 和 wi_baryi 累积的中间结果，用 $Sum_of_wi_bar$ 和 $Sum_of_wi_baryi$ 分别表示 wi_bar 和 wi_baryi 累积的最终结果；用 $V_hat_y_bar_hat$ 表示 $\hat{V}(\hat{y})$ ；用 $FRRRC$ 和 RRC 分别表示采用固定效应模型和随机效应模型时的合并效应 RR；用 $FRRLOW$ 、 $FRRUP$ 及 $RRLOW$ 、 $RRUP$ 分别表示采用固定效应模型和随机效应模型时合并效应 RR 的 95% CI 的下限和上限。

设分析例 14-6 定性资料的程序名为 SASTJFX14_6.SAS。


```

OPTIONS LS =74 PS =MAX CENTER NODATE;
DATA A;
INPUT study_ID n_ai n_bi n_ci n_di @@ ;
n_ai +0.25;n_bi +0.25;n_ci +0.25;n_
di +0.25;
OR = (n_ai* n_di)/(n_bi* n_ci);
yi =LOG(OR);
wi = (1/n_ai + 1/n_bi + 1/n_ci + 1/n_
di)** (-1);
Ni =n_ai +n_bi +n_ci +n_di;
number_of_study =9;
CARDS;
1 63 92 29 126
2 103 95 113 112
3 25 218 17 195
4 18 162 25 252
5 27 70 52 207
6 14 70 15 126
7 29 52 16 61
8 122 89 92 113
9 49 55 30 74
;
RUN;
DATA B;SET A;
wi2=wi** 2;wiyi =wi* yi;wiyi2 =wi
* yi** 2;
sum_of_wi_temp +wi;sum_of_wi2_temp
+wi2;
sum_of_wiyi_temp +wiyi;
sum_of_wiyi2_temp +wiyi2;
ID =_N_;
IF _N_ =number_of_study THEN DO;
sum_of_wi =sum_of_wi_temp;
sum_of_wi2 =sum_of_wi2_temp;
sum_of_wiyi =sum_of_wiyi_temp;
sum_of_wiyi2 =sum_of_wiyi2_temp;
yw_bar =sum_of_wiyi/sum_of_wi;
square_of_s_y_bar =1/sum_of_wi;
Q =sum_of_wiyi2 -yw_bar** 2* sum_of
_wi;
DF =number_of_study -1;
P =1 - PROBCHI(Q,DF);
I2 = (Q - (number_of_study - 1))/Q
* 100;

IF I2 <0 THEN I2 =0;
FORC =EXP(yw_bar);
FORLOW =EXP(yw_bar -1.96* square_of
_s_y_bar** 0.5);
FORUP =EXP(yw_bar +1.96* square_of
_s_y_bar** 0.5);
w_bar =sum_of_wi/number_of_study;
Square_of_s_w = (Sum_of_wi2 - number
_of_study* w_bar** 2)/(number_of
_study -1);
U = (number_of_study - 1)* (w_bar -
Square_of_s_w/(number_of_study* w_
bar));
Square_of_tao_hat = (Q - (number_of
_study -1))/U;
IF Square_of_tao_hat <0 THEN Square_
of_tao_hat =0;
END; RUN;
PROC SORT;BY DESCENDING ID;
DATA C;SET B;
Square_of_tao_hat_for_all + Square_
of_tao_hat;
wi_bar =1/(1/wi + Square_of_tao_hat_
for_all);
RUN;
PROC SORT;BY ID;
DATA D;SET C;
Sum_of_wi_bar_temp +wi_bar;
Sum_of_wi_baryi_temp +wi_bar* yi;
IF _N_ =number_of_study THEN DO;
Sum_of_wi_bar =Sum_of_wi_bar_temp;
Sum_of_wi_baryi =Sum_of_wi_baryi_
temp;
V_hat_y_bar_hat =1/Sum_of_wi_bar;
y_bar_hat =Sum_of_wi_baryi/Sum_of
_wi_bar;
ORC =EXP(y_bar_hat);
ORLOW =EXP(y_bar_hat -1.96* V_hat_y
_bar_hat** 0.5);
ORUP =EXP(y_bar_hat +1.96* V_hat_y
_bar_hat** 0.5);
END;RUN;
PROC SORT;BY DESCENDING ID;
DATA E;SET D;WHERE ID = number_of_
study;RUN;
PROC PRINT;
VAR number_of_study Q DF P I2 FORC
FORLOW FORUP ORC ORLOW ORUP;
ODS HTML;RUN;

```

程序中分别用 n_{ai} 、 n_{bi} 、 n_{ci} 、 n_{di} 表示各研究对应四格表 4 个格子上的频数；用 N_i 表示各研究的样本含量；用 y_i 表示优势比 OR 的对数值；用 w_i 表示采用固定效应模型时的权重；用 w_i^2 表示 w_i 的平方；用 $w_i y_i$ 表示 w_i 与 y_i 的乘积；用 $w_i y_i^2$ 表示 w_i 与 y_i 的平方的乘积；用

number_of_study 表示纳入研究的数量；用 Sum_of_wi_temp、Sum_of_wi2_temp、Sum_of_widi_temp 和 Sum_of_widi2_temp 分别表示 wi、wi2、widi 和 widi2 累积的中间结果，用 Sum_of_wi、Sum_of_wi2、Sum_of_widi 和 Sum_of_widi2 分别表示 wi、wi2、widi 和 widi2 的累积的最终结果；用 yw_bar 和 y_bar_hat 分别表示采用固定效应模型和随机效应模型时的合并效应；用 w_bar、Square_of_s_w、Square_of_tao_hat 和 wi_bar 分别表示 \bar{w} 、 s_w^2 、 $\hat{\tau}^2$ 和 \bar{w}_i ；用 DF 表示自由度；用 Q 和 P 分别表示异质性检验的 χ^2 值和 P 值；I2 为异质性检验的另一个统计量；用 Sum_of_wi_bar_temp 和 Sum_of_wi_baryi_temp 分别表示 wi_bar 和 wi_baryi 累积的中间结果，用 Sum_of_wi_bar 和 Sum_of_wi_baryi 分别表示 wi_bar 和 wi_baryi 累积的最终结果；用 V_hat_y_bar_hat 表示 $\hat{V}(\hat{y})$ ；用 FORC 和 ORC 分别表示采用固定效应模型和随机效应模型时的合并效应 OR；用 FORLOW、FORUP 及 ORLOW、ORUP 分别表示采用固定效应模型和随机效应模型时合并效应 OR 的 95% CI 的下限和上限。

设分析例 14-7 定性资料的程序名为 SASTJFX14_7.SAS。

```
DATA SROC_INITIAL;
INPUT ID nTP nFP nFN nTN @@ ;
N=nTP+nFP+nFN+nTN;
TPR=(nTP+0.5)/(nTP+0.5+nFN+0.5);
FPR=(nFP+0.5)/(nFP+0.5+nTN+0.5);
FNR=(nFN+0.5)/(nFN+0.5+nTP+0.5);
TNR=(nTN+0.5)/(nTN+0.5+nFP+0.5);
CARDS;
1 30 2 3 55
2 22 2 0 25
3 56 3 3 89
4 53 11 7 132
5 30 15 14 95
6 4 0 2 18
7 9 3 7 28
8 39 7 4 25
9 7 1 4 8
10 18 1 0 23
;
RUN;
DATA SROC_VALID;SET SROC_INITIAL;
WHERE FPR<0.5;
V=LOG((nTP+0.5)/(nFN+0.5));
U=LOG((nFP+0.5)/(nTN+0.5));
S=V+U;D=V-U;RUN;
proc iml;reset print;
use SROC_VALID(drop = ID N TPR FPR
FNR TNR V U S D);
read all into matrix_data;
sum_c=matrix_data[+,];
NP_total=sum_c[1,1]+sum_c[1,3];
NN_total=sum_c[1,2]+sum_c[1,4];
matrix_tempN=NP_total||NN_total;
create data_tempN var{NP_total NN
_total};
append from matrix_tempN;quit;

TPR=(1+((1-FPR)/FPR)**((1+B)/(
(1-B))*EXP(-A/(1-B))))**(-1);
TPR=ROUND(TPR,0.01);
Se=TPR;Sp=1-FPR;FNR=1-TPR;
TNR=1-FPR;
Distance=SQRT((TPR-1)**2+(
FPR)**2);
OUTPUT;END;RUN;
proc iml;reset print;
use A_B(keep =TPR FPR FNR TNR
Distance);
read all into matrix_Distance;
pM=matrix_Distance[>:,5];
MinimumDistance=matrix_Distance
[pM,];
create data_MinimumDistance
var {FPR TPR FNR TNR MinimumDis-
tance};
append from MinimumDistance;quit;
DATA MinimumDistance;
MERGE data_MinimumDistance data_
tempN;Se=TPR;Sp=TNR;
Sse=SQRT(Se*(1-Se)/NP_total);
Ssp=SQRT(Sp*(1-Sp)/NN_total);
SeLOW=Se-1.96*Sse;SeUP=Se+1.
96*Sse;
SpLOW=Sp-1.96*Ssp;SpUP=Sp+1.
96*Ssp;RUN;
DATA SCATTER(KEEP =TPRi FPRi);
SET SROC_INITIAL;
RENAME TPR =TPRi FPR =FPRi;RUN;
DATA GRAPH;SET SCATTER A_B;
MERGE SCATTER A_B;RUN;
GOPTION RESET =ALL HSIZE =6 VSIZE =
5 ftExt = '宋体';
SYMBOL I =SPLINE VALUE =DOT;
```

```

PROC REG DATA = SROC_VALID
OUTEST = AB;
MODEL D = S; RUN;
DATA A_B; SET AB; A = INTERCEPT;
B = S; DO
FPR = 0.01, 0.02, 0.05, 0.06, 0.07,
0.08, 0.09, 0.10, 0.11, 0.12, 0.13,
0.14, 0.15, 0.16, 0.17, 0.18, 0.19,
0.20, 0.30, 0.40, 0.50, 0.60, 0.70,
0.80, 0.90;
AXIS1 LABEL = ('误诊率') ORDER = (0.0
TO 1.0 BY 0.1) MINOR = NONE OFFSET =
(0,0);
AXIS2 LABEL = (ANGLE = 90 '灵敏度')
ORDER = (0.0 TO 1.0 BY 0.1) MINOR
= none OFFSET = (0,0);
LEGEND1 ACROSS = 1 CBORDER = black
LABEL = none VALUE = ('合并' '单项')
POSITION = (top right inside);
PROC GPLOT DATA = GRAPH;
PLOT TPR* FPR TPRi* FPRi = 'CIRCLE' /
HAXIS = AXIS1 legend = legend1 VAXIS
= AXIS2 OVERLAY;
RUN;
PROC PRINT DATA = MinimumDistance;
ODS HTML; RUN;

```

DATA 步进行原始数据录入,接着校正原始数据,PROC REG 步计算回归系数和截距,然后计算某些工作点对应的指标值,并绘制 SROC 曲线。

14.3.5 主要分析结果及解释

例 14-5 的输出结果:

| Obs | number_ of_study | Q | DF | P | I ² |
|---------|---------------------|---------|---------|---------|----------------|
| 1 | 7 | 9.90697 | 6 | 0.12862 | 39.4365 |
| FRRC | FRRLOW | FRRUP | RRC | RRLOW | RRUP |
| 0.91449 | 0.86656 | 0.96507 | 0.89345 | 0.80136 | 0.99612 |

异质性检验的结果为: $Q = \chi^2 = 9.90697$, $df = 7 - 1 = 6$, $P = 0.12862 > 0.05$, $I^2 = 39.4365\% < 50\%$, 因此认为各研究同质性较好,采用固定效应模型进行综合分析是合适的。结果表明:平均效应尺度(RR)约为 0.91,其 95% 置信区间为 0.87 ~ 0.97。由于置信区间位于 1 的左侧,所以试验组与对照组的死亡危险率之间的差别具有统计学意义,7 项研究的综合结果显示,使用阿司匹林预防心肌梗死后死亡具有一定的效果。使用阿司匹林预防后,可以使死亡的危险降低约 9 个百分点。

例 14-6 的输出结果:

| Obs | number_of_study | Q | DF | P | I ² |
|---------|-----------------|---------|---------|---------|----------------|
| 1 | 9 | 13.0864 | 8 | 0.10891 | 38.8679 |
| FORC | FORLOW | FORUP | ORC | ORLOW | ORUP |
| 1.60059 | 1.34273 | 1.90797 | 1.63207 | 1.29358 | 2.05912 |

异质性检验的结果为: $Q = \chi^2 = 13.0864$, $df = 9 - 1 = 8$, $P = 0.10891 > 0.05$, $I^2 = 38.8679\% < 50\%$, 因此认为各研究同质性较好,采用固定效应模型进行综合分析是合适的。结果表明:平均效应尺度(OR)约为 1.60,其 95% 置信区间约为 1.34 ~ 1.91。由于置信区间位于 1 的右侧,

故得出“病例组与对照组的暴露比值之间的差别具有统计学意义”之统计学结论，以及“高暴露于低频电磁场是儿童患白血病的危险因素”之专业结论。

例 14-7 的输出结果如图 14-1 所示。

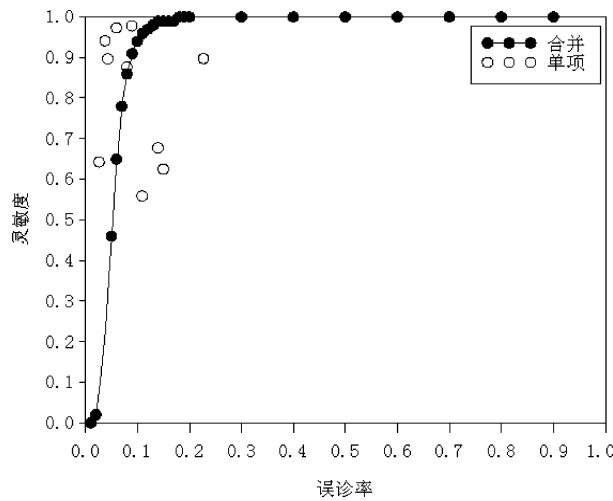


图 14-1 根据 10 项研究估计的 SROC 曲线

| Obs | FPR | TPR | FNR | TNR | MINIMUMDISTANCE | NP_TOTAL | NN_TOTAL |
|------|------|----------|----------|---------|-----------------|----------|----------|
| 1 | 0.1 | 0.94 | 0.06 | 0.9 | 0.11662 | 312 | 543 |
| Se | Sp | Sse | Ssp | SeLOW | SeUP | SpLOW | SpUP |
| 0.94 | 0.90 | 0.013445 | 0.012874 | 0.91365 | 0.96635 | 0.87477 | 0.92523 |

这里给出的是 SROC 曲线及曲线上距离左上角最近的点，即诊断指标正常/异常值的最佳临界点对应的各统计指标的值，其中的灵敏度和特异度为该诊断试验的综合灵敏度和综合特异度，而且两者都比较高：Se = 0.94，Sp = 0.90，其 95% CI 分别为 (0.91, 0.97) 和 (0.87, 0.93)。

14.4 用 ROC 方法分析诊断试验资料

14.4.1 问题与数据

【例 14-8】 利用宏程序绘制 ROC 曲线并计算曲线下面积。假定有如下一组资料，涉及“是否患病”和“某定量指标”两个变量，分别用“disease”和“age”表示，前者为定性变量，后者为定量变量。研究者想用“某定量指标”来判断受试者是否患该种疾病。试利用宏程序绘制 ROC 曲线并计算曲线下面积。数据见相应的 SAS 程序，注意，程序中的 bw 变量代表与此项研究关系不大的另一个定量指标(本例中未利用此变量所对应的数据)。

【例 14-9】 对例 14-8 资料，试检验 ROC 曲线下面积与 0.5 的差别有无统计学意义。

【例 14-10】 利用宏程序比较 ROC 曲线下面积。采用某种金标准，将 45 例患者中的 21 例确诊为 RMSF(病例组)，其余 24 例确诊为非 RMSF(对照组)，分别用两种方法测得血钠水平如表 14-8 所示。问：这两种方法诊断性能之间的差异有无统计学意义？

表 14-8 45 例患者的血钠水平

| 金标准 诊断结果 | 血钠水平 (mmol/L) | | | | | | | | | | | |
|-------------|---------------|-----|-----|-----|-----|-----|------|-----|-----|-----|-----|-----|
| | 方法 1 | | | | | | 方法 2 | | | | | |
| RMSF 病例 | 124 | 125 | 126 | 126 | 127 | 128 | 122 | 124 | 125 | 125 | 126 | 126 |
| | 128 | 128 | 128 | 129 | 129 | 131 | 127 | 128 | 128 | 128 | 130 | 130 |
| | 132 | 133 | 133 | 135 | 135 | 135 | 133 | 133 | 134 | 134 | 134 | 134 |
| | 136 | 138 | 139 | | | | 136 | 138 | 140 | | | |
| 非 RMSF 病例 | 129 | 131 | 131 | 134 | 134 | 135 | 124 | 128 | 130 | 133 | 133 | 133 |
| | 136 | 136 | 136 | 137 | 137 | 138 | 134 | 134 | 134 | 134 | 136 | 136 |
| | 138 | 139 | 139 | 139 | 139 | 140 | 137 | 138 | 138 | 140 | 140 | 141 |
| | 140 | 141 | 142 | 142 | 142 | 143 | 141 | 142 | 142 | 142 | 142 | 144 |

14.4.2 对数据结构的分析

由例 14-8 可以看出, 该资料实际涉及一个原因变量(定量观测指标)和一个定性结果变量(两种测定方法), 一个二值结果变量(是否患某病)。

由例 14-10 可以看出, 研究者分别用两种方法测定 45 例患者的血钠水平, 根据患者是否具有 RMSF 特征, 将他们分为病例组与对照组, 因此, 该资料应属于具有一个重复测量的两因素设计一元定量资料。其中, 实验分组因素是诊断结果(RMSF 病例与非 RMSF 病例), 重复测量因素为“检测方法”, 定量的结果变量为“血钠水平”。但在诊断试验中, 希望将血钠水平视为定量的原因变量, 而把诊断结果视为结果变量。

14.4.3 分析目的与统计分析方法的选择

例 14-8 和例 14-10 的研究目的都是希望找到定量指标的合适取值(即分割点或诊断界值), 基于此可对受检者做出明确地诊断, 并使诊断结果准确度高, 最好是特异度和灵敏度同时都高。

在临床医学诊断试验中, 如果能采用金标准来诊断疾病是最为理想的, 但有时受到客观条件的限制, 无法利用金标准来诊断。目前, ROC 分析已成为描述和比较诊断试验精确度的标准之一。对角线上方的 ROC 曲线下面积被认为具有合理的辨别能力, 可用来诊断受试者是否患有所要诊断的疾病。

14.4.4 SAS 程序中重要内容的说明

设分析例 14-8 定性资料(指结果变量 disease, 而原因变量 age 是定量变量, 计算时, 将 age 由小到大排序, 并依次用每一个 age 的数值作为一个“假定的诊断点”, 与金标准一道构造出一系列的四格表, 从而, 可以计算出 ROC 曲线上的各个点对应的坐标, 并依据这些点绘制出 ROC 曲线)的程序名为 SASTJFX14_8.SAS。

```
data SASTJFX14_8;
input disease age bw @@ ;
datalines;
0 50 65 0 39 61 0 21 70 0 61 67
0 30 55 0 35 63 0 25 72 0 41 66
0 43 52 0 36 54 0 37 76 0 25 61
%include "D:\SASTJFX\rocplot.sas";
ods html;
ods graphics on;
proc logistic data = SASTJFX14_9;
model disease(event = "1") = age/
```

```

0 41 53 0 62 55 0 28 70 0 33 68
1 52 45 1 49 61 1 47 42 1 62 31
1 55 67 1 70 61 1 75 55 1 77 52
1 81 58 1 64 53 1 62 41 1 39 57
1 61 51 1 61 55 1 57 49 1 79 47
;
run;

```

```

outroc = roc1 roceps = 0;
output out = outp p = phat;
run;
%rocplot(outroc = roc1, out = outp,
         p = phat, id = age);
ods graphics off;
ods html close;

```

在 SAS 官方网页上提供了“绘制 ROC 曲线”(http://support.sas.com/kb/25/addl/fusion250114_4_rocplot.sas.txt)和“用非参数方法比较相关 ROC 曲线下面积”(http://support.sas.com/kb/25/addl/fusion_25017_6_roc.sas.txt)的两个宏程序。前者主要介绍如何快速绘制出比 SAS 说明文档中更加美观实用的 ROC 曲线。后者主要介绍如何检验 ROC 曲线下面积与 0.5 的差异有无统计学差异。假定这两个程序 roc.sas 和 rocplot.sas 分别保存在 C 盘的 ROC 文件夹中。

%include 宏命令, 可以将 rocplot.sas 程序调入到本 SAS 程序中; 在本程序中, 将 disease 选项中的“1”视为患有疾病, 并将绘制 ROC 曲线所需的数据输出到 roc1 数据集中, 将每个观测值年龄的预测概率数据输出到 outp 数据集中, 并将预测概率命名为 phat; 对 rocplot 宏程序, outroc 指定的数据集要与 proc logistic 中的 outroc 数据集相同; out 指定的数据集要与 proc logistic 中的 output 中的 out 数据集相同; p 指定的预测概率的名称要与 output 中 p 所命名的名称相同; id 标记 ROC 曲线上的原始数值。

设分析例 14-9 定性资料的程序名为 SASTJFX14_9.SAS。

```

data SASTJFX14_9;
INPUT G $ NUMS;
DO I = 1 TO NUMS;
INPUT VALUE @@ ;
OUTPUT; END;
CARDS;
ABNORMAL 16
52 49 47 62
55 70 75 77
81 64 62 39
61 61 57 79
NORMAL 16
50 39 21 61
30 35 25 41
43 36 37 25
41 62 28 33
;
PROC FREQ ORDER = FORMATTED
PAGE;
TABLES G* VALUE /NOPRINT
SPARSE OUT = A ;
DATA B;
SET A NOBS = KK; K = KK/2;
PUT K = ;
IF G = 'NORMAL' THEN XNN = COUNT;
ELSE XAA = COUNT;
PROC MEANS NWAY NOPRINT;

```

```

DATA F;
SET F;
ARRAY XN(*) XN1 - XN&K; ARRAY XA(*) XA1 - XA&K;
ARRAY NN1(&K);          ARRAY NA1(&K);
ARRAY YN(&K);           ARRAY YA(&K);
ARRAY AREAS(&K);
ARRAY Q1S(&K);          ARRAY Q2S(&K);
NN = SUM(OF XN1 - XN&K);
NA = SUM(OF XA1 - XA&K);
DO I = 1 TO &K;          DO J = 1 TO I;
NN1(I) + XN(J);          NA1(I) + XA(J);
YN(I) + LAG(XN(J));      /* YA(I) + LAG(XA(J)); */
END;
YA(I) = NA - NA1(I); /* YN(I) = NN - NN1(I); */
IF YN(I) = . THEN YN(I) = 0;
/* IF YA(I) = . THEN YA(I) = 0; */
AREAS(I) = XN(I) * YA(I) + 1/2 * XN(I) * XA(I);
Q1S(I) = XN(I) * (YA(I) ** 2 + XA(I) * YA(I) + 1/
3 * XA(I) ** 2);
Q2S(I) = XA(I) * (YN(I) ** 2 + XN(I) * YN(I) + 1/
3 * XN(I) ** 2);
AREA = SUM(OF AREAS1 - AREAS&K) / (NN * NA);
Q1 = SUM(OF Q1S1 - Q1S&K) / (NN * NA ** 2);
Q2 = SUM(OF Q2S1 - Q2S&K) / (NA * NN ** 2);
END;
SE_AREA = SQRT((AREA * (1 - AREA) + (NA - 1) *
(Q1 - AREA ** 2) +

```

```

VAR XNN XAA; CLASS VALUE;
OUTPUT OUT=C SUM=S1 S2 ;
PROC TRANSPOSE DATA=C OUT=D
PREFIX=XN; VAR S1;
PROC TRANSPOSE DATA=C OUT=
E PREFIX=XA; VAR S2;
DATA F;
SET D; SET E;
RUN;
% LET K=25;

(NN-1)*(Q2-AREA**2)/(NA*NN));
Z=(AREA-0.5)/SE_AREA;
P=(1-PROBNORM(Z))*2;
LCL95_A=AREA-PROBIT(0.975)*SE_AREA;
UCL95_A=AREA+PROBIT(0.975)*SE_AREA;
AREA=ROUND(AREA,0.0001);
SE_AREA=ROUND(SE_AREA,0.0001);
Z=ROUND(Z,0.0001);
IF P<0.0001 THEN P=0.0001;
P=ROUND(P,0.0001);
LCL95_A=ROUND(LCL95_A,0.0001);
UCL95_A=ROUND(UCL95_A,0.0001);
FILE PRINT;
PUT #2 @5'ROC 曲线下面积 AZ='AREA
/@5'面积的标准误 SE(AZ)='SE_AREA
/@5'面积与 0.5 差异的假设检验统计量 Z='Z
/@5'假设检验对应的 P 值='P
/@5'ROC 曲线下面积的 95% 置信区间为 ('
LCL95_A','UCL95_A');
RUN;

```

对于连续性资料 ROC 曲线的计算,也可理解为对诊断界值很多的有序分类资料进行计算,此时每一个不重复的测量值都要作为诊断界值考虑,程序思路与有序分类资料相似,但需要先算出不重复的测量值个数或诊断分类个数。需将数据部分中两组的 16 及 16 个数据删掉,换为自己的数据,并将表示各组数量的“16”和“16”换为自己各组数据的个数。首先运行一遍程序,此时不必看结果,而要先从程序记录窗口中读出 k 值,即不重复的测量值个数,然后重新调回刚发送的程序,程序中的“% LET K=65;”中的“65”表示不重复的测量值个数,将此数值替换掉,重新运行程序,即可得到正确的结果。如果实际测量值越大,结果异常的概率越小,则还要修改程序中含有“/*”和“*/”的三行语句,具体把三行中左侧的语句用“/*”和“*/”屏蔽掉,而把此三行中右侧的“/*”和“*/”删掉。

设分析例 14-10 定性资料的程序名为 SASTJFX14_10.SAS。

```

data SASTJFX14_10;
input Method1 Method2 disease @@ ;
cards;
124 122 1 125 124 1 126 125 1
126 125 1 127 126 1 128 126 1
128 127 1 128 128 1 128 128 1
129 128 1 129 130 1 131 130 1
132 133 1 133 133 1 133 134 1
135 134 1 135 134 1 135 134 1
136 136 1 138 138 1 139 140 1
129 124 0 131 128 0 131 130 0
134 133 0 134 133 0 135 133 0
136 134 0 136 134 0 136 134 0
137 134 0 137 136 0 138 136 0
138 137 0 139 138 0 139 138 0
139 140 0 139 140 0 140 141 0
;

%roc(data=outp_Method1 outp_Method2,
var=phat_Method1 phat_Method2,
response=disease);
symbol1 i=join v=circle c=blue line=1;
symbol2 i=join v=dot c=green line=2;
ods html close;
data roc1_Method1;
set roc1_Method1;
index="Method1";
run;
data roc1_Method2;
set roc1_Method2;
index="Method2";
run;

```

```

140 141 0   141 142 0   142 142 0
142 142 0   142 142 0   143 144 0
;
run;
%include "c:\roc\roc.sas";
ods html
file="D:\SASTJFX\SASTJFX14_10.htm"
style=statdoc;
proc logistic data=SASTJFX14_10;
model disease(event="1")=Method1/
outroc=roc1_Method1 roceps=0;
output out=outp_Method1 p=phat_Method1;
run;
proc logistic data=SASTJFX14_10;
model disease(event="1")=Method2/
outroc=roc1_Method2 roceps=0;
output out=outp_Method2 p=phat_Method2;
run;

data joint;
set roc1_Method1 roc1_Method2;
run;
axis1 order=(0 to 1 by .1);
axis2 order=(0 to 1 by .1) label=(angle
=90);
filename grafout
"D:\SASTJFX\SASTJFX14_10.gif";
goptions device=gif
gsfname=grafout
gsfmode=replace
ftext=;
proc gplot data=joint;
label index="Index";
plot_sensit_*_1mspec_=index /haxis
=axis1 vaxis=axis2;
run;
quit;

```

本程序的特别说明：调用了两次 logistic 回归分析过程，第1个 model 语句等号右边写上了“Method1”，其含义是用方法一检测血钠值，并让血钠值由小到大的每一个血钠值充当“假定的诊断点”，与金标准一道构造出一系列的四格表，从而计算出一系列的坐标点 $[(1 - \text{特异度}), \text{灵敏度}]$ ，用折线将这些点中相邻的两点连接起来，便可绘制出一条 ROC 曲线；同理，可以得出与“方法二”对应的另一条 ROC 曲线。若将这两个过程步之一的 model 语句等号右边的“Method1”或“Model2”删掉，就能达到前面例 14-8 和例 14-9 的分析目的，即计算出一条 ROC 曲线下面积，并将此面积与 0.5 进行比较。

14.4.5 主要分析结果及解释

例 14-8 的输出结果：

该输出结果由两部分组成：第一部分见下面文本格式输出的内容；第二部分见图 14-2。

| Association of Predicted Probabilities and Observed Responses | | | |
|--|------|----------|-------|
| Percent Concordant | 90.2 | Somers'D | 0.824 |
| Percent Discordant | 7.8 | Gamma | 0.841 |
| Percent Tied | 2.0 | Tau - a | 0.425 |
| Pairs | 256 | c | 0.912 |

ROC 曲线下面积正好等于统计量 c 的值，即 0.912，由此可以判定诊断准确性较高。但是，ROC 曲线下面积 0.912 与 0.5 的差别有无统计学意义，还需要做进一步假设检验。

（说明：此输出结果存放位置在“D:\SASTJFX”文件夹内，文件名分别为 SASTJFX14_10.htm 和 SASTJFX14_10.gif。）

例 14-9 的输出结果：

ROC 曲线下面积 $AZ = 0.9121$

面积的标准误 $SE(AZ) = 0.0506$

面积与 0.5 差异的假设检验统计量 $Z = 8.1378$

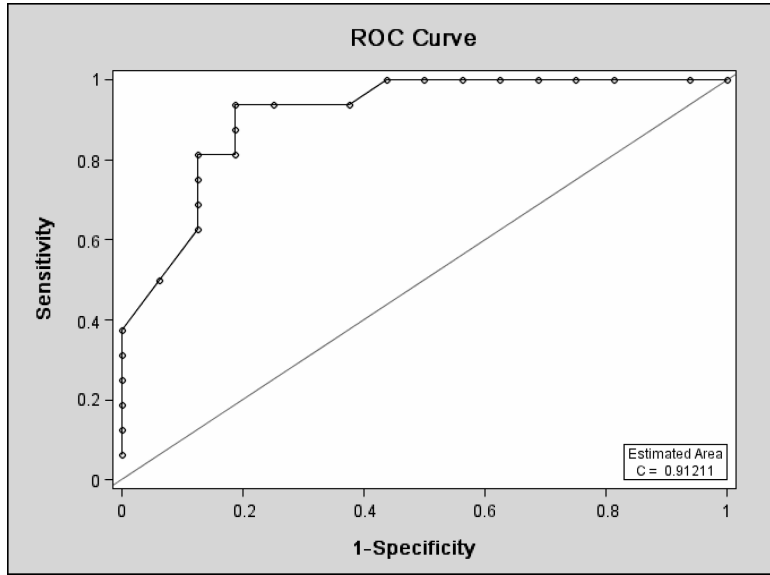


图 14-2 用宏程序绘制的 ROC 曲线

假设检验对应的 P 值 = 0.0001

ROC 曲线下面积的 95% 置信区间为 (0.8129, 1.0114)。

专业结论: 因 ROC 曲线下面积为 0.9121, 此值与无效假设下的面积 0.5 之间的差别有统计学意义 ($Z = 8.1378$, $P = 0.0001$), 表明用“给定的定量指标”来判断受试者是否患该种疾病的诊断方法, 诊断效果是令人满意的。

例 14-10 的输出结果:

该输出结果由两部分组成: 第一部分见下面文本格式输出的内容; 第二部分见图 14-3。

为了区别由 Method1 和 Method2 这两种方法产生的结果, 这里将 Method1 在 output 中的输出结果定义为 outp_Method1, 预测概率定义为 phat_Method1; 将 Method2 在 output 中的输出结果定义为 outp_Method2, 预测概率定义为 phat_Method2。

ROC Curve Areas and 95% Confidence Intervals

| | ROC Area | Std Error | Confidence | Limits |
|--------------|----------|-----------|------------|--------|
| phat_Method1 | 0.8750 | 0.0505 | 0.7760 | 0.9740 |
| phat_Method2 | 0.8075 | 0.0643 | 0.6815 | 0.9336 |

Tests and 95% Confidence Intervals for Contrast Rows

| | Estimate | Std Error | Confidence | Limits | Chi-square | Pr > ChiSq |
|------|----------|-----------|------------|--------|------------|------------|
| Row1 | 0.0675 | 0.0213 | 0.0257 | 0.1092 | 10.0307 | 0.0015 |

Contrast Test Results

| Chi-Square | DF | Pr > ChiSq |
|------------|----|------------|
| 10.0307 | 1 | 0.0015 |

在“ROC Curve Areas and 95% Confidence Intervals”的表格中, 我们可以看到 Method1 的 ROC 曲线下面积为 0.875, 而 Method2 的 ROC 曲线下面积为 0.8075。二者相比较的 P 值为 0.0015 (< 0.05), 所以这两条 ROC 曲线下面积差异有统计学意义。

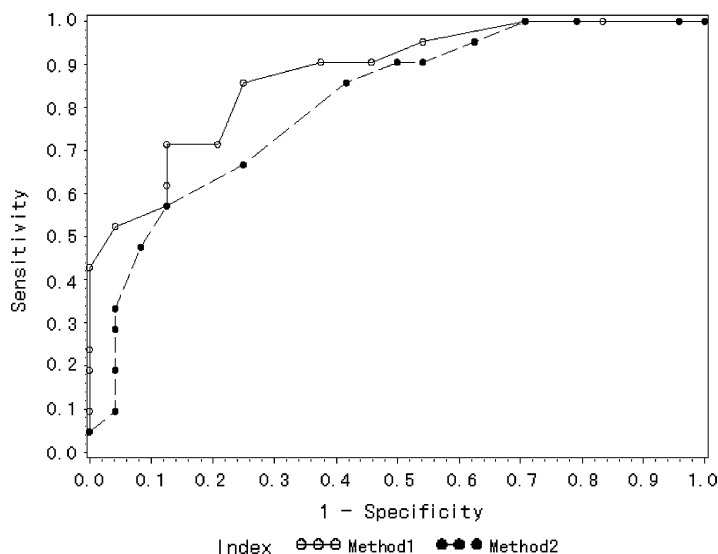


图 14-3 两种方法对应的 ROC 曲线

专业结论为：两种方法的诊断性能之间是有差异的。方法 1 和方法 2 对应的 ROC 曲线下面积分别为 0.875 和 0.8075，前者比后者大，方法 1 的诊断性能好于方法 2。

14.5 本章小结

本章针对高维列联表资料，给出了差异性分析方法。其中，很容易混淆的两种统计分析方法分别是加权 χ^2 检验和 CMH χ^2 检验。前者通常适用于结果变量为二值变量的高维列联表资料，其分析目的是消除多个原因变量的影响，考察剩下的一个原因变量与结果变量之间是否独立；而后者可用于结果变量具有三种性质（即二值、多值有序和多值名义）的高维列联表资料，因为它包含了三种假设检验。当结果变量为二值变量的高维列联表资料时，CMH χ^2 检验中的“一般关联性分析”的假设检验的作用与加权 χ^2 检验类似，但有区别，它给出 OR 或 RR 的估计值及关于 OR 或 RR 是否等于 1 的假设检验。本章还介绍了定性资料的 meta 分析方法、ROC 曲线分析方法及 SAS 实现方法。

（柳伟伟 周诗国 曹波 王琪）

第 15 章 多因素设计一元定性资料的对数线性模型分析

本章主要介绍利用 SAS 软件实现多因素设计一元定性资料的对数线性模型分析。多因素设计一元定性资料通常以高维列联表的形式呈现，对高维列联表采取压缩降维或分层分析的办法，这样做有时是有效的，但如果将某重要的、与保留因素不独立的因素压缩后进行分析，易导致因素效应混杂，容易得到错误的结果。当同时分析多个变量的主效应和变量间的交互效应时，对数线性模型分析便显示出其巨大的潜力。

15.1 问题、数据及统计分析方法的选择

15.1.1 问题与数据

【例 15-1】 为研究影响新生儿健康状况的有关影响因素，某研究者调查了 5251 名产妇与新生儿出生情况的资料，包括产妇年龄、是否吸烟(含被动吸烟)、怀孕前是否服用某种避孕药及新生儿健康情况(正常、异常)四个变量，变量名分别简记为 A、S、D、Y。数据见表 15-1。

表 15-1 5251 例新生儿健康状况与产妇因素的关系

| 年龄(A) | 吸烟(S) | 服药(D) | 例数(人) | |
|-------|-------|-------|------------|-----|
| | | | 婴儿健康状况(Y): | |
| <30 | 是 | 是 | 204 | 58 |
| | | 否 | 330 | 67 |
| | 否 | 是 | 1051 | 210 |
| | | 否 | 1014 | 178 |
| 30~34 | 是 | 是 | 125 | 31 |
| | | 否 | 180 | 42 |
| | 否 | 是 | 582 | 144 |
| | | 否 | 489 | 85 |
| 35~ | 是 | 是 | 35 | 20 |
| | | 否 | 35 | 10 |
| | 否 | 是 | 158 | 53 |
| | | 否 | 119 | 31 |

【例 15-2】 某研究单位对 63 名肺炎患儿进行两项生理指标的检测，结果见表 15-2，其中有 32 名结核性、13 名化脓性和 18 名细菌性肺炎。请分析这两种指标与患儿肺炎类型之间的关系。

表 15-2 63 例肺炎患儿的检测结果

| 指标一 | 指标二 | 例数(人) | | | |
|-----|-----|-------|-----|-----|-----|
| | | 肺炎类型: | 结核性 | 化脓性 | 细菌性 |
| 阴性 | 阴性 | | 4 | 5 | 4 |
| | 阳性 | | 9 | 1 | 1 |
| 阳性 | 阴性 | | 9 | 6 | 11 |
| | 阳性 | | 10 | 1 | 2 |

15.1.2 对数据结构的分析

例 15-1 资料涉及三个原因变量,即产妇年龄、是否吸烟、怀孕前是否服用某种避孕药;结果变量为新生儿健康情况,其取值为正常或异常,属于二值变量,故该资料属于多因素设计一元定性资料。

例 15-2 资料共涉及三个定性变量,分别为两个定性的原因变量(指标一和指标二)与一个定性的结果变量(肺炎类型)。肺炎类型的取值有三个水平,属于多值名义变量,故该资料属于多因素设计一元定性资料。

15.1.3 分析目的与统计分析方法的选择

对例 15-1 而言,该资料若是进行差异性分析,可以采用 CMH χ^2 检验;若是进行证实性研究,可以新生儿健康情况(Y)为因变量,以产妇年龄、吸烟与服药为自变量,拟合 logistic 回归模型;若是探索性研究,除分析年龄、吸烟与否、是否服药对新生儿健康状况的影响外,还要考察各个自变量间的关系以及各自变量对新生儿健康的影响是否有联合作用,可采用对数线性模型分析。

对数线性模型是指一族模型,主要用于离散型数据的分析。对于复杂高维列联表的分析,采用交叉乘积比分析高维交互作用,模型通过交互效应项反映各变量间的关系及其效应大小。

对例 15-2 而言,结果变量为多值名义变量的高维列联表资料可以选用的统计分析方法有 CMH χ^2 检验、扩展的多重 logistic 回归分析和对数线性模型。这里采用对数线性模型进行分析,其原理与分析结果变量为二值变量的高维列联表资料基本相同。

15.2 用对数线性模型分析列联表资料

15.2.1 对数线性模型简介

对数线性模型是分析高维列联表行之有效的办法,最先由 Yule、Bartlett 利用 Yule(1900 年)定义的交叉乘积比分析三维交互作用,然后由 Kullback(1968 年)引入方差分析的思想发展而来。

对数线性模型把各分组变量(包括自变量和因变量)水平组合下期望(理论频数)的自然对数表示为各分组变量及其交互作用的线性函数,通过迭代计算求得模型中参数的估计值,进而运用方差分析的思想检验各主效应和交互作用的效应大小。

15.2.2 用 SAS 分析例 15-1 资料

1. SAS 程序及其说明

(1) 饱和模型

饱和模型是指对数线性模型中独立参数的个数等于列联表的网格数,即模型中包含所有因素的主效应和各级交互效应。设对例 15-1 中资料拟合饱和模型的程序名为 SASTJFX15_1A. SAS, 程序如下:

| | |
|--|--|
| <pre> Data; input Y A S D wt@@ ; cards; 1 1 1 1 204 1 1 1 2 330 1 1 2 1 1051 1 1 2 2 1014 1 2 1 1 125 1 2 1 2 180 1 2 2 1 582 1 2 2 2 489 1 3 1 1 35 1 3 1 2 35 1 3 2 1 158 1 3 2 2 119 2 1 1 1 58 2 1 1 2 67 2 1 2 1 210 2 1 2 2 178 2 2 1 1 31 2 2 1 2 42 2 2 2 1 144 2 2 2 2 85 2 3 1 1 20 2 3 1 2 10 2 3 2 1 53 2 3 2 2 31 ; run; </pre> | <pre> ODS HTML STYLE = MINIMAL; proc catmod data = ; weight wt; model A* S* D* Y = _response_; loglin A S D Y; run; ODS HTML CLOSE; </pre> |
|--|--|

程序说明：首先建立数据集，变量 wt 表示频数，其他变量的含义可参见例题中的叙述。语句“proc catmod data = prg15_1;”表示调用 CATMOD 过程；“model A * S * D * Y = _response_;”语句中等号左端的 A * S * D * Y 指明要分析的变量，等号右端的 _response_ 表示拟合对数线性模型；“loglin A | S | D | Y;”语句中的 A | S | D | Y 表示拟合所有的主效应与交互效应，即饱和模型，该语句等同于“loglin A S D Y A * S A * D A * Y S * D S * Y D * Y A * S * D A * S * Y A * D * Y S * D * Y A * S * D * Y”。

(2) 二阶交互效应模型

该模型包括主效应和二阶及二阶以下的各阶交互效应，实现该分析的程序名为 SAS-TJFX15_1B. SAS。与饱和模型程序的不同在于“loglin A | S | D | Y @ 3;”语句，该语句表示拟合二阶交互效应模型。由于只有 loglin 语句与饱和模型的程序略有不同，这里不再给出具体程序，后面的一阶交互效应模型和最优模型的程序同样如此。

(3) 一阶交互效应模型

该模型包括主效应和所有的一阶交互效应，实现该分析的程序名为 SASTJFX15_1C. SAS。“loglin A | S | D | Y @ 2;”语句表示拟合一阶交互效应模型。

(4) 最优模型的筛选

在实际应用中，需要拟合各种分层或谱系模型，找出既能拟合样本资料又有较大自由度的“最优模型”。随着列联表维度增加，谱系模型数量剧增。如何从众多的谱系模型中找出最优的模型是件困难而又有实际意义的工作。最优模型筛选通常采用类似逐步回归方法中的前进法或后退法，简单易行，一般的统计软件也以此来筛选模型。为避免忽略有意义的高阶效应，建议采用后退法。

CATMOD 过程并不像多重回归分析那样给出最优模型的自动筛选过程，好在一般列联表维度不一定太高，而且遵循一定模型筛选规则，获取最优模型并不是遥不可及。一般而言，我们可以首先拟合几个一致阶模型，找出那些没有统计意义的交互效应项。然后从高阶模型向低阶模型，逐步剔除那些无统计学意义的交互项。当有多个可能的最优模型或交互效应项 P 值接近 0.05 时，可进一步借助 AIC 等评价模型拟合优度的统计量，做出综合评判选择。

本例中拟合最优模型的程序名为 SASTJFX15_1D. SAS。语句“loglin A S D Y A * D A * Y S * D S * Y D * Y;”中包含了主效应和除 A * S 以外的其他一阶交互效应。

2. 主要分析结果及解释

(1) 饱和模型

Maximum Likelihood Analysis of Variance

| Source | DF | Chi-Square | Pr > ChiSq |
|------------------|----|------------|------------|
| A | 2 | 654.74 | < .0001 |
| S | 1 | 513.56 | < .0001 |
| A * S | 2 | 0.46 | 0.7964 |
| D | 1 | 3.02 | 0.0825 |
| A * D | 2 | 11.69 | 0.0029 |
| S * D | 1 | 12.92 | 0.0003 |
| A * S * D | 2 | 4.67 | 0.0969 |
| Y | 1 | 648.16 | < .0001 |
| A * Y | 2 | 11.81 | 0.0027 |
| S * Y | 1 | 4.83 | 0.0280 |
| A * S * Y | 2 | 0.41 | 0.8141 |
| D * Y | 1 | 7.99 | 0.0047 |
| A * D * Y | 2 | 0.83 | 0.6601 |
| S * D * Y | 1 | 0.30 | 0.5822 |
| A * S * D * Y | 2 | 2.29 | 0.3189 |
| Likelihood Ratio | 0 | . | . |

“Maximum Likelihood Analysis of Variance”是方差分析表，使用最大似然法进行分析。这里四个因素之间的三阶交互作用 A * S * D * Y 经检验证明没有统计学意义，三个因素之间的二阶交互作用也是没有统计学意义的，一阶交互作用中 A * S 没有统计学意义。由于这里建立的是饱和模型，已经没有剩余的自由度分配给似然比检验，所以并没有关于模型拟合情况的似然比检验的结果。

Analysis of Maximum Likelihood Estimates

| Parameter | | Estimate | Standard Error | Chi-Square | Pr > ChiSq |
|-----------|-------|----------|----------------|------------|------------|
| A | 1 | 0.7817 | 0.0316 | 612.54 | < .0001 |
| | 2 | 0.2035 | 0.0347 | 34.38 | < .0001 |
| S | 1 | -0.6100 | 0.0269 | 513.56 | < .0001 |
| | 2 | -0.0182 | 0.0316 | 0.33 | 0.5651 |
| A * S | 1 1 | -0.0182 | 0.0316 | 0.33 | 0.5651 |
| | 2 1 | 0.0127 | 0.0347 | 0.13 | 0.7134 |
| D | 1 | 0.0467 | 0.0269 | 3.02 | 0.0825 |
| | 2 | -0.0997 | 0.0316 | 9.97 | 0.0016 |
| A * D | 1 1 | -0.0997 | 0.0316 | 9.97 | 0.0016 |
| | 2 1 | -0.0426 | 0.0347 | 1.51 | 0.2193 |
| S * D | 1 1 | -0.0968 | 0.0269 | 12.92 | 0.0003 |
| | 2 1 | -0.00652 | 0.0316 | 0.04 | 0.8364 |
| A * S * D | 1 1 1 | -0.00652 | 0.0316 | 0.04 | 0.8364 |
| | 2 1 1 | -0.0744 | 0.0347 | 4.60 | 0.0320 |
| Y | 1 | 0.6853 | 0.0269 | 648.16 | < .0001 |
| | 2 | 0.0900 | 0.0316 | 8.11 | 0.0044 |
| A * Y | 1 1 | 0.0900 | 0.0316 | 8.11 | 0.0044 |
| | 2 1 | 0.0642 | 0.0347 | 3.42 | 0.0645 |
| S * Y | 1 1 | -0.0592 | 0.0269 | 4.83 | 0.0280 |
| | 2 1 | 0.0642 | 0.0347 | 3.42 | 0.0645 |

| | | | | | |
|---------------|---------|----------|--------|------|--------|
| A * S * Y | 1 1 1 | -0.00311 | 0.0316 | 0.01 | 0.9217 |
| | 2 1 1 | 0.0221 | 0.0347 | 0.40 | 0.5247 |
| D * Y | 1 1 | -0.0761 | 0.0269 | 7.99 | 0.0047 |
| A * D * Y | 1 1 1 | 0.0178 | 0.0316 | 0.32 | 0.5727 |
| | 2 1 1 | 0.0243 | 0.0347 | 0.49 | 0.4831 |
| S * D * Y | 1 1 1 | -0.0148 | 0.0269 | 0.30 | 0.5822 |
| A * S * D * Y | 1 1 1 1 | -0.0111 | 0.0316 | 0.12 | 0.7254 |
| | 2 1 1 1 | 0.0513 | 0.0347 | 2.19 | 0.1392 |

“Analysis of Maximum Likelihood Estimates”给出了模型中参数的估计值以及对其进行假设检验的结果。这部分结果与方差分析表的结果一致，对于三阶及二阶交互作用对应参数的检验基本上都是没有统计学意义的。参数估计值默认将每个效应的最后一类作为参照类，其他各个水平通过与参照类相比来分析效应大小。如年龄与新生儿健康情况的交互效应(A * Y)在两个变量取值都为 1 时的参数估计值为 0.0900(P = 0.0044)，是一正值，表示：低年龄孕妇组(A = 1)较高年龄孕妇组(A = 3)，有更大可能生育健康的新生儿(正常：Y = 1)。

(2) 二阶交互效应模型

Maximum Likelihood Analysis of Variance

| Source | DF | Chi-Square | Pr > ChiSq |
|------------------|----|------------|------------|
| A | 2 | 664.09 | < .0001 |
| S | 1 | 528.30 | < .0001 |
| A * S | 2 | 0.51 | 0.7739 |
| D | 1 | 3.08 | 0.0795 |
| A * D | 2 | 12.68 | 0.0018 |
| S * D | 1 | 14.46 | 0.0001 |
| A * S * D | 2 | 2.72 | 0.2563 |
| Y | 1 | 659.91 | < .0001 |
| A * Y | 2 | 12.69 | 0.0018 |
| S * Y | 1 | 5.69 | 0.0171 |
| A * S * Y | 2 | 0.50 | 0.7777 |
| D * Y | 1 | 8.19 | 0.0042 |
| A * D * Y | 2 | 0.78 | 0.6768 |
| S * D * Y | 1 | 0.17 | 0.6812 |
| Likelihood Ratio | 2 | 2.30 | 0.3168 |

这部分仅给出方差分析的结果，三个因素之间的二阶交互作用仍然是没有统计学意义的。似然比检验的结果为 $\chi^2 = 2.30$, P = 0.3168。

(3) 一阶交互效应模型

Maximum Likelihood Analysis of Variance

| Source | DF | Chi-Square | Pr > ChiSq |
|--------|----|------------|------------|
| A | 2 | 709.67 | < .0001 |
| S | 1 | 558.52 | < .0001 |
| A * S | 2 | 1.67 | 0.4338 |
| D | 1 | 1.45 | 0.2291 |
| A * D | 2 | 14.43 | 0.0007 |

| | | | |
|------------------|---|--------|--------|
| S * D | 1 | 51.53 | <.0001 |
| Y | 1 | 743.92 | <.0001 |
| A * Y | 2 | 17.06 | 0.0002 |
| S * Y | 1 | 7.03 | 0.0080 |
| D * Y | 1 | 10.10 | 0.0015 |
| Likelihood Ratio | 9 | 6.88 | 0.6492 |

一阶交互效应模型也仅给出方差分析的结果。A * Y、S * Y、D * Y 的交互效应均有统计学意义，提示产妇年龄、吸烟、服药都是新生儿异常的影响因素，A * S 仍无统计学意义，表明吸烟与年龄并无关联。似然比检验的结果为 $\chi^2 = 6.88$, $P = 0.6492$ 。

(4) 最优模型的筛选

Maximum Likelihood Analysis of Variance

| Source | DF | Chi-Square | Pr > ChiSq |
|------------------|----|------------|------------|
| A | 2 | 940.63 | <.0001 |
| S | 1 | 822.27 | <.0001 |
| D | 1 | 1.43 | 0.2312 |
| Y | 1 | 741.81 | <.0001 |
| A * D | 2 | 13.86 | 0.0010 |
| A * Y | 2 | 17.26 | 0.0002 |
| S * D | 1 | 50.97 | <.0001 |
| S * Y | 1 | 7.22 | 0.0072 |
| D * Y | 1 | 10.11 | 0.0015 |
| Likelihood Ratio | 11 | 8.55 | 0.6636 |

由于在上述拟合过程中，A * S 无统计学意义，我们将之剔除出模型。最终的模型包括的一阶交互作用有 A * D、A * Y、S * D、S * Y、D * Y，方差分析的结果表明这5项的作用都是有统计学意义的。似然比检验的结果 $\chi^2 = 8.55$, $P = 0.6636 > 0.05$ ，说明模型对资料的拟合较好，可以认为该模型是成立的。

Analysis of Maximum Likelihood Estimates

| Parameter | | Estimate | Standard Error | Chi-Square | Pr > ChiSq |
|-----------|-----|----------|----------------|------------|------------|
| A | 1 | 0.7872 | 0.0265 | 879.85 | <.0001 |
| | 2 | 0.2008 | 0.0291 | 47.59 | <.0001 |
| S | 1 | -0.6129 | 0.0214 | 822.27 | <.0001 |
| D | 1 | 0.0284 | 0.0237 | 1.43 | 0.2312 |
| Y | 1 | 0.6827 | 0.0251 | 741.81 | <.0001 |
| A * D | 1 1 | -0.0797 | 0.0214 | 13.86 | 0.0002 |
| | 2 1 | -0.00866 | 0.0234 | 0.14 | 0.7118 |
| A * Y | 1 1 | 0.0984 | 0.0264 | 13.86 | 0.0002 |
| | 2 1 | 0.0491 | 0.0290 | 2.88 | 0.0899 |
| S * D | 1 1 | -0.1213 | 0.0170 | 50.97 | <.0001 |
| S * Y | 1 1 | -0.0574 | 0.0214 | 7.22 | 0.0072 |
| D * Y | 1 1 | -0.0583 | 0.0183 | 10.11 | 0.0015 |

由参数估计值的符号可进一步获知该效应是正向还是反向的,是保护效应还是危险效应。如效应 $A * Y$ 两个估计值均为正,表示相对于 35 岁以上产妇而言,年龄偏小的产妇更可能生育正常婴儿;而 $S * Y$ 、 $D * Y$ 符号为负,表明吸烟与否(S)与服用避孕药与否(D)为危险因素,即表示相对于不吸烟、未服用避孕药的产妇,那些吸烟与服用避孕药的产妇更有可能导致新生儿异常。同时,相对于 logistic 回归分析而言,还可以获得自变量之间关联性的信息,如 $A * D$ 表明低年龄组较少服用避孕药; $S * D$ 表明吸烟或被动吸烟与怀孕前是否服用避孕药呈现反向关联;而未引入 $A * S$ 项时,分析者认为“吸烟与年龄无关联”。

专业结论:结合实际资料可知,产妇年龄较小,新生儿异常的比例较低;而吸烟与服用避孕药会导致新生儿异常情况的增加。

15.2.3 用 SAS 分析例 15-2 资料

1. SAS 程序及其说明

建立饱和模型后发现,三个因素的二阶交互作用无统计学意义,所以应该寻找简单一些的模型。考察所有可能的不饱和模型,模型的选择应该依据拟合优度检验的结果,在模型成立的基础上,要求模型尽量简单。在进行综合比较后,最后发现当只有指标二和肺炎类型一个交互作用项纳入时,模型比较合适。本文只讨论这种情形下的 SAS 程序及输出结果,其他步骤可以参考例 15-1 的做法。

对例 15-2 中资料拟合对数线性模型的程序名为 SASTJFX15_2. SAS, 程序如下:

```
data abc;
  do x1 = 0 to 1;
    do x2 = 0 to 1;
      do type = 1 to 3;
        input f @@ ;
        output;
      end;
    end;
  end;
cards;
4 5 4 9 1 1 9 6 11 10 1 2
;
run;
```

```
ODS HTML STYLE = MINIMAL;
proc catmod data = abc;
  weight f;
  model x1* x2* type = _response_;
  loglin x2 type x2 * type;
run;
ODS HTML CLOSE;
```

程序说明: 首先建立数据集 abc, 用变量 x1、x2 和 type 分别表示指标一、指标二和肺炎类型。过程步中的语句与例 15-1 基本相同, 其中“loglin x2 type x2 * type;”语句表示模型中只包含 x2 * type 这一个交互效应。

2. 主要分析结果及解释

Maximum Likelihood Analysis of Variance

| Source | DF | Chi-Square | Pr > ChiSq |
|------------------|----|------------|------------|
| x2 | 1 | 7.69 | 0.0056 |
| type | 2 | 11.29 | 0.0035 |
| x2 * type | 2 | 11.18 | 0.0037 |
| Likelihood Ratio | 6 | 5.85 | 0.4397 |

方差分析的结果显示, 对指标二、肺炎类型和它们的交互作用项进行检验, 三者都具有统

计学意义。似然比检验的结果 $\chi^2 = 5.85$, $P = 0.4397 > 0.05$, 说明模型对资料的拟合较好, 可以认为该模型是成立的。

| Analysis of Maximum Likelihood Estimates | | | | | |
|--|-----|----------|----------------|------------|------------|
| Parameter | | Estimate | Standard Error | Chi-Square | Pr > ChiSq |
| x2 | 0 | 0.4891 | 0.1764 | 7.69 | 0.0056 |
| | 1 | 0.6868 | 0.2047 | 11.25 | 0.0008 |
| type | 2 | -0.5223 | 0.2835 | 3.39 | 0.0654 |
| | 0 1 | -0.6789 | 0.2047 | 10.99 | 0.0009 |
| x2 * type | 0 2 | 0.3633 | 0.2835 | 1.64 | 0.2000 |

在参数估计的结果中, 指标二有两个水平, 所以参数估计部分对应一个结果; 肺炎类型有三个水平, 因此参数估计的结果有两个; 这两个因素交互作用项的参数估计结果也有两个。这部分结果与方差分析的结果是一致的。

专业结论: 由于指标二和肺炎类型的交互作用项有统计学意义, 同时结合实际资料可知, 在指标二阴性和阳性的患儿中, 三种类型肺炎患儿的构成比不同, 在指标二为阳性的患儿中, 结核性肺炎的患儿所占比例要明显大于指标二为阴性的患儿。

15.3 本章小结

本章介绍了多因素设计一元定性资料的对数线性模型分析, 这种设计类型的资料常常以高维列联表的形式呈现, 当然, 也可以是原始的数据库形式。在分析时, 首先应该正确辨别数据结构, 然后根据分析的目的合理地选择分析方法。在本章介绍的两个例子中, 既包括结果变量是二值变量的情形, 也包括了结果变量是多值名义变量的情形。对数线性模型分析在 SAS 软件中通过 CATMOD 过程来完成, 这里仅介绍了一些最基本的用法, 其他细节和一些特殊的情形可参阅 SAS 说明书。

(张岩波 柳伟伟 李长平)

第 5 篇 对定量结果进行预测性分析

第 16 章 两变量简单线性相关与回归分析

在医学科学研究中，常常要分析变量之间的关系，以说明事物发生、发展及变化的原因或变量间依存变化的数量关系，如人的年龄与血压、血脂，毒物剂量与动物的死亡率，环境中污染物的浓度与污染源的距离的关系等。要分析变量间存在的关系，可采用回归和相关分析方法。如果分析的变量是两个，则通常采用简单线性相关与回归分析方法；若分析的变量超过两个，则应采用多重回归分析方法。本章仅针对简单相关与回归分析问题，介绍在 SAS 软件中如何用 CORR 过程进行简单相关分析、用 REG 过程进行简单回归分析，以及用 PROBIT 过程进行半数效量估计的加权直线回归分析。

16.1 问题、数据及统计分析方法的选择

16.1.1 问题与数据

【例 16-1】 某医学临床研究为探讨正常人血液中钾元素含量与心肌中钾元素含量是否存在相关关系，测定了 20 例意外死亡的正常人血液中钾元素含量与心肌中钾元素平均含量，测定结果如表 16-1 所示。试进行相关分析。

【例 16-2】 某研究测量不同病理分级前列腺癌患者的血清 T-PSA，30 例前列腺癌患者中低分化癌(Ⅰ)10 例，中分化癌(Ⅱ)10 例，高分化癌(Ⅲ)10 例，各分化组的平均血清 T-PSA 值如表 16-2 所示。试分析病理分级与血清 T-PSA 值间有无相关性。

表 16-1 正常人血液中钾元素含量 (mmol/L)
与心肌中钾元素平均含量 (mg/g)

| 编号 | 血液中钾含量 X (mmol/L) | 心肌中钾平均含量 Y (mg/g) |
|-----|----------------------|----------------------|
| 1 | 9.88 | 66.37 |
| 2 | 10.72 | 73.34 |
| ... | ... | ... |
| 20 | 8.74 | 67.99 |

表 16-2 不同分化程度前列腺癌的
血清 T-PSA 水平 (ng/mL)

| 患者编号 | 分化水平 | T-PSA (ng/mL) |
|------|------|---------------|
| 1 | Ⅰ | 127.9 |
| 2 | Ⅰ | 69.0 |
| ... | ... | ... |
| 30 | Ⅲ | 61.1 |

【例 16-3】 用快速法和 ELISA 法对同一批样品进行抗体检测试验，结果见表 16-3。问：这两种检测方法的检测结果之间是否具有相关性？

表 16-3 两种测定方法同时检测抗体得 73 个样品的检测结果

| 快速法检测结果 | 样品数 | | | | 合 计 | |
|---------|----------|----|----|----|-----|-----|
| | ELISA 法: | - | + | ++ | | +++ |
| - | | 15 | 0 | 2 | 3 | 20 |
| + | | 2 | 19 | 1 | 2 | 24 |
| ++ | | 1 | 3 | 17 | 0 | 21 |
| +++ | | 0 | 2 | 0 | 6 | 8 |
| 合计 | | 18 | 24 | 20 | 11 | 73 |

【例 16-4】 土壤内 NaCl 含量对植物的生长有很大影响，NaCl 含量过高，将增加组织内无机盐累积，抑制植物生长。表 16-4 给出每千克土壤中 NaCl 的含量(X)、植物单位叶面积干物重量(Y)。试进行简单线性回归分析。

表 16-4 不同土壤 NaCl 含量对植物单位叶面积干物重量的影响

| | | | | | | | |
|-----------------------------------|----|-----|-----|-----|-----|-----|-----|
| 土壤 NaCl 含量 X/g · kg ⁻¹ | 0 | 0.8 | 1.6 | 2.4 | 3.2 | 4.0 | 4.8 |
| 干重 Y/mg · dm ⁻² | 80 | 90 | 95 | 115 | 130 | 115 | 135 |

【例 16-5】 在突变实验中，用不同剂量的射线，照射植物的种子，发现苗期高度与成活株之间有一定的关系。用 X 射线照射大麦的种子，记处理株第一叶平均高度占对照株高度的百分数为 X，存活百分数为 Y，得到如表 16-5 所示结果。试分析苗期高度是否影响成活株数量。

表 16-5 大麦处理株第一叶平均高度占对照株高度的百分数与存活百分数

| | | | | | | | | | | |
|---|----|----|----|----|----|----|----|----|----|-----|
| X | 40 | 41 | 41 | 43 | 47 | 51 | 52 | 55 | 61 | 62 |
| Y | 60 | 61 | 67 | 68 | 70 | 85 | 88 | 89 | 90 | 100 |

注：数据为分析需要虚构，非真实数据。

【例 16-6】 某人以 1 种已知的毒物(标号为 1)作为对照，来研究另 2 种未知毒物(标号分别为 2 和 3)的毒性大小，每种毒物均用了若干个剂量，每个剂量下分别用若干只大鼠做了实验。设毒物分组标志为 A，剂量为 DOSE，各次实验的死亡数为 R、实验动物数为 N，资料格式见表 16-6，详细数据参见完整的 SAS 程序。试计算各种毒物的半数致死量 LD₅₀(mg/kg)，并把 2 种未知毒物分别与对照毒物相比较。

表 16-6 毒物种类与剂量对大鼠死亡与否的影响结果

| 毒物种类 | 剂量 (mg/kg) | R: 死亡数 | n: 各剂量组大鼠数 |
|------|------------|--------|------------|
| 1 | 0.3 | 0 | 8 |
| 1 | 0.4 | 2 | 8 |
| ... | ... | ... | ... |
| 3 | 1.0 | 9 | 10 |

16.1.2 对数据结构的分析

在例 16-1 中，整个资料只涉及一个组，即测定了 20 例意外死亡的正常人血液中钾元素含量与心肌中钾元素平均含量，故应该属于“单组设计”。指标(血液中钾元素含量和心肌中钾平均含量)为测量得到且可以带小数，故该资料属于定量资料，即这是单组设计定量资料。

在例 16-2 中，资料只有一个组，即测定了 30 例不同病理分级前列腺癌患者的血清 T-PSA 水平，故应该属于“单组设计”。指标(T-PSA 水平)为测量得到且可以带小数，故该资料属于定量资料，即这是单组设计定量资料。

在例 16-3 中，该资料属于双向有序且属性相同的方表资料，也可称为配对设计扩大形式的定性资料。

在例 16-4 中,该资料只涉及一个组,即测定了 7 种不同土壤 NaCl 含量下植物单位叶面积干物的重量,故应该属于“单组设计”。指标“干重”为测量得到且可以带小数,故该资料属于定量资料,即这是单组设计定量资料。

在例 16-5 中,该资料只涉及一个组,即不同剂量的射线照射植物的种子,测量处理株第一叶平均高度占对照株高度的百分数和存活百分数,该资料属于定量资料,即这是单组设计定量资料。

在例 16-6 中,该资料涉及 3 个组,分别是对照组、毒物 2 和毒物 3,每种毒物均用了若干个剂量,并且每种毒物下的各个剂量都不相同,每个剂量下分别用了若干只大鼠做了实验,观测指标为“各次实验的动物是死还是活(即二值结果变量)”,若不知“毒物种类”与“剂量”两个因素中的哪一个对结果变量的影响作用大,则该资料为两因素析因设计定性资料。若把每种毒物所对应的特定剂量组中的全部动物视为一个整体,计算出死亡率,用其作为新的结果变量,则该资料转化为无重复实验的两因素设计定量资料。

16.1.3 分析目的与统计分析方法的选择

对例 16-1 而言,该研究的分析目的是欲考察正常人血液中钾元素含量与心肌中钾元素含量是否存在相关关系。本例中, X 、 Y 都是随机变量,且各指标数据的分布基本服从正态分布,故对该资料进行相关分析是可行的。首先绘制散点图,看两变量间是否呈线性变化趋势,如果有线性趋势则进行下一步相关分析,并对相关系数进行假设检验。

对例 16-2 而言,该研究分析的目的是分析病理分级与血清 T-PSA 值间有无相关性。当两变量不符合双变量正态分布时,需要用 Spearman 秩相关分析来考察变量间的关系,该资料的病理分级属于有序变量,血清 T-PSA 水平属于定量的随机变量,因此分析该资料应该使用 Spearman 秩相关分析。

对例 16-3 而言,通常的分析目的为两种检测方法检测的结果之间是否具有一致性,应选用 Kappa 检验;另一个分析目的是考察两种方法检测结果之间是否具有关联性。由于每个样品都用两种方法来检测,所以,它属于“自身配对设计扩大的形式”,测自同一个样品的两个数据之间具有较高的相关性,此时,进行关联性分析不太适合采用 Spearman 秩相关分析(适用于单因素两水平设计定量资料或有序资料的秩相关分析),而宜计算 Kendall's Tau-b 相关系数并进行相应的假设检验。

对例 16-4 而言,该研究分析的目的是想考察不同土壤 NaCl 含量对植物单位叶面积干物重量的影响,该资料中土壤 NaCl 含量(X)和植物单位叶面积干物重量(Y)都是随机变量,且各指标具体数据基本服从正态分布,故对该资料进行简单相关与回归分析是可行的。首先应对资料绘制散点图,看两变量间是否呈线性变化趋势,如果有线性趋势再进行下一步的线性相关和回归分析。

对例 16-5 而言,该研究资料的分析目的是想考察苗期高度是否影响成活株数量。本例中处理株第一叶平均高度占对照株高度的百分数(X)和存活百分数(Y)都是随机变量,且各指标具体数据的分布基本服从正态分布,故对该资料进行相关与回归分析是可行的。首先应对该资料绘制散点图,看两变量间是否呈线性变化趋势,如果有线性趋势再进行下一步的直线相关和回归分析。

对例 16-6 而言,该研究的分析目的有两个:一是计算各种毒物的半数致死量 LD_{50} (mg/kg),二是把 2 种未知毒物分别与对照毒物相比较。要实现目的一需要采用半数效量估计的加权直

线回归分析方法。由于加权直线回归分析过程不能直接实现目的二，所以需要采用 SAS 编程和统计公式来间接实现。

16.1.4 统计分析方法简介

分析例 16-1 资料将用到 Pearson 直线相关分析，有关概念简介如下：当两个变量取值之间出现一个增大、另一个也增大(或减小)的情况时，我们称这种现象为共变，也就是说这两个变量之间有“相关关系”。简单线性相关分析是描述两定量变量间是否有直线关系以及直线关系的方向和密切程度的分析方法。此分析方法主要是通过计算相关系数的大小并对其进行假设检验以及结合专业知识来评价得到的相关系数是否有实际意义来完成的。

分析例 16-2 资料将用到 Spearman 秩相关分析，有关概念简介如下：在做 Pearson 相关分析时，要求两变量服从正态分布，然而若得到的原始数据并不服从正态分布或其总体分布未知，有时数据中存在所谓“超限值”(如限于仪器的灵敏度，仅知道血样中某物质浓度小于 $0.001\mu\text{g/mL}$)，甚至数据本身就是等级资料，此时宜采用等级相关或称秩相关来分析两变量的线性联系程度与方向。这类方法是利用两变量的秩次大小进行线性相关分析，对原变量分布不做要求，属非参数统计分析方法。

分析例 16-3 资料需要采用 Kendall's Tau-b 相关分析方法。所谓 Kendall's Tau-b 相关分析，就是基于 Kendall 提出的计算公式，算出反映配对设计扩大情形下两有序变量间呈现的关联度高低和方向的 Kendall's Tau-b 相关系数，简记为 τ ，并依据样本相关系数 τ ，推断其所代表的总体相关系数 ρ_τ 是否为 0。与 Pearson 乘积 - 矩相关分析和 Spearman 秩相关分析有所不同，其统计推断仅包括假设检验。

分析例 16-4 和例 16-5 资料将用到简单线性回归分析，有关概念简介如下：简单线性回归分析是用直线回归方程表示两个定量变量间依存关系的统计分析方法，此分析方法主要由三部分组成，其一，计算反映两定量变量依赖关系的直线回归方程，即计算直线回归方程的截距 a 、斜率 b ；其二，根据样本截距 a 、斜率 b ，检验样本所抽自的总体截距 α 是否为 0、总体斜率 β 是否为 0；其三，结合专业知识，评价此直线回归方程是否有实用价值。

分析例 16-6 资料将用到加权线性回归分析，有关概念简介如下：半数数量(ED_{50})是实验物质引起实验动物总体中半数产生某种反应所需的剂量，通常以 mg/kg 表示。若剂量用浓度(mg/L)或时间作为标志，则称半数有效浓度(EC_{50})或半数有效时间(ET_{50})；若反应用死亡、耐受或抑制作为标志，则称半数致死量(LD_{50})、半数耐受量(ELM_{50})或半数抑制量(ID_{50})。其中， LD_{50} 用得最多，它在药理学及毒理学研究中应用甚广。

当绘出剂量反应曲线(剂量为横轴，死亡率为纵轴)，会发现曲线呈长尾 S 形，如将剂量取对数后，则剂量反应曲线呈对称的 S 形。此曲线两端伸延较缓，说明在低剂量与高剂量区域内，剂量即使变化较大，但引起反应率的变化却很小，而在曲线中段，斜率较大，特别在死亡率 $p=50\%$ 处剂量稍有改变，就会引起反应率的明显变化，说明 LD_{50} 甚为敏感，故选择半数数量作为评价指标，对鉴别不同药物或毒物的毒性大小，具有较高的敏感性。

由于研究曲线的规律比较困难，人们发现：将反应率转换成概率单位后，便将对称的 S 形曲线直线化了，这给研究半数数量带来了极大的方便。

把反应率转化成概率单位的方法是：①直接查百分数 p 与概率单位对照表；②如果没有此对照表，把反应率看成正态曲线下面积，根据面积，反查“标准正态曲线下面积表”，得到标准正态变量的 u 值，用查得的各 u 值加 5，便得到与各反应率 P 相对应的概率单位值。

半数效量的概率单位法是多种计算半数效量方法中最有效的一种，最先由 C. I. Bliss 提出，故简称为 Bliss 法。由于概率单位是非正态的，且方差不齐，故不适合用通常的最小二乘法直接拟合概率单位随对数剂量变化的直线回归方程，需用各点上方差的倒数做权重，进行加权，并用最大似然法 (Maximum Likelihood Method) 求解找出一个回归模型的参数估计值，故此法又称为概率单位法或最大似然法。

16.2 Pearson 线性相关性分析

16.2.1 SAS 程序中重要内容的说明

程序名为 SASTJFX16_1.SAS。

```
data SASTJFX16_1;input x y;cards;
  9.88  66.37
 10.72  73.34
 10.36  70.79
  ...  ...
;
ods html;
```

```
proc gplot;plot y*x;run;
proc corrFISHER (alpha = 0.05 bias-
adj=no) pearson;
var x y;
run;
ods html close;
```

数据步，输入两变量 x 、 y ，依次录入数据。过程步，“proc gplot; plot $y * x$; run;”语句是绘制变量 y 关于 x 的散点图，“proc corr FISHER (alpha = 0.05 biasadj = no) pearson; var $x y$; run;”语句是对变量 y 与 x 进行相关性分析，“FISHER (alpha = 0.05 biasadj = no)”选项是分析变量 y 与 x 之间总体相关系数的 95% 置信区间，并且不对偏移进行校正。

16.2.2 主要分析结果及解释

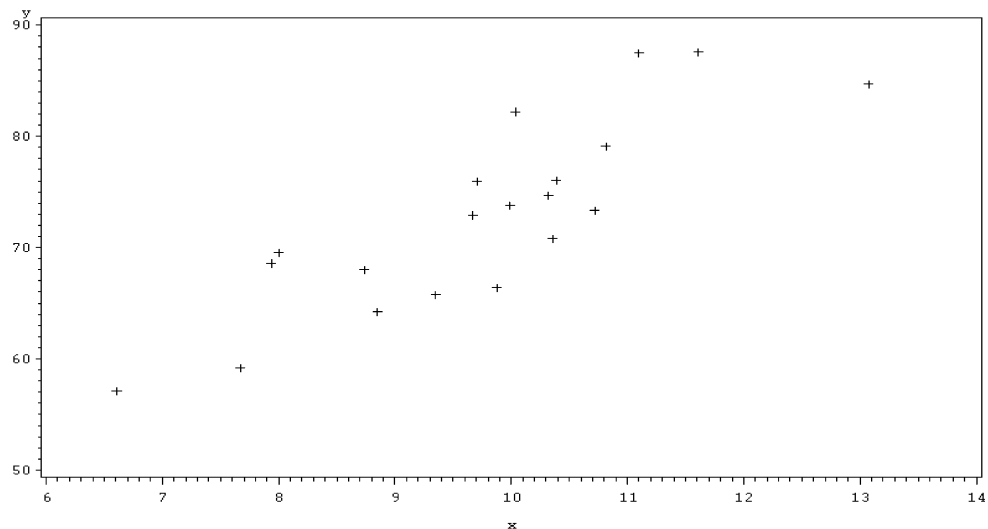


图 16-1 变量 y 与 x 的散点图

由图 16-1 的散点图可见，20 个点在一条长带内随机分布着，且不存在明显的曲线趋势，说明两个变量之间存在一定的直线相关关系，可进行直线相关性分析。

Pearson 相关系数, N = 20

当 $H_0: \rho = 0$ 时, $\text{Prob} > |r|$

| | x | y |
|---|---------|---------|
| x | 1.00000 | 0.84636 |
| | | <.0001 |
| y | 0.84636 | 1.0000 |
| | <.0001 | |

Pearson 相关统计量; 基于 Fisher 变换

| 变量 | 带变量 | N | 样本相关 | Fisher's z | 95% 置信限 | p Value for $H_0: \rho = 0$ |
|----|-----|----|---------|------------|-------------------|-----------------------------|
| x | y | 20 | 0.84636 | 1.24319 | 0.645665 0.937688 | <.0001 |

在直线相关分析结果中, 给出了 Pearson 相关系数 r 值以及假设检验的结果, $r = 0.84636$, $P < 0.0001$, 并给出了总体相关系数的 95% 置信区间为 $(0.645665, 0.937688)$, 相关系数不包含 0, 即相关系数与 0 的差异有统计学意义, 与假设检验的结果一致, 说明变量 y 与 x 之间存在的直线相关关系有统计学意义, 即血液中钾含量与心肌中钾平均含量之间存在相关关系。

16.3 Spearman 秩相关分析

16.3.1 SAS 程序中重要内容的说明

程序名为 SASTJFX16_2.SAS。

```
data SASTJFX16_2; input x y @@ ;
cards;
1 127.9 1 69.0
.....
3 68.9 3 25.1
;
run;
```

```
ods html; proc corr spearman;
var x y;
run;
ods html close;
```

数据步, 输入两变量 x 、 y , 依次录入数据。过程步, `proc corr` 调用相关分析过程, `spearman` 选项用于实现 Spearman 秩相关分析。

16.3.2 主要分析结果及解释

Pearson 相关系数, N = 30

当 $H_0: \rho = 0$ 时, $\text{Prob} > |r|$

| | x | y |
|---|----------|----------|
| x | 1.00000 | -0.58015 |
| | | 0.0008 |
| y | -0.58015 | 1.0000 |
| | 0.0008 | |

以上为 Spearman 秩相关分析结果, $r = -0.58015$, $P = 0.0008 < 0.05$, 说明变量 x 和 y 之间呈负相关关系, 即随着癌分化水平的加重, 血清 T-PSA 值降低。

16.4 Kendall's Tau-b 相关分析

16.4.1 SAS 程序中重要内容的说明

程序名为 SASTJFX16_3.SAS。

```
DATA SASTJFX16_3;
  DO A=1 TO 4;
    DO B=1 TO 4;
      INPUT F @@ ;
      OUTPUT;
    END;
  END;
CARDS;
15 0 2 3
2 19 1 2
1 3 17 0
0 2 0 6
;
RUN;
ODS HTML;
PROC CORR KENDALL;
  VAR A B;
  FREQ F;
RUN;
ODS HTML CLOSE;
```

【程序说明】 在“PROC CORR”语句中，不要加用于求总体相关系数置信区间的选项“FISHER”，SAS 系统认为此时不适合求其置信区间。

16.4.2 主要分析结果及解释

| Kendall Tau b 相关系数, N = 73 | | |
|--|--------------------|--------------------|
| 当 $H_0: \rho = 0$ 时, $\text{Prob} > r $ | | |
| | A | B |
| A | 1.00000 | 0.56680 < .0001 |
| B | 0.56680 < .0001 | 1.00000 |

求得 Kendall' Tau-b 相关系数为 $\tau = 0.56680$, $P < 0.0001$ 。说明总体相关系数与 0 之间的差别具有统计学意义。

【统计与专业结论】 因 $\tau = 0.56680$, $P < 0.0001$, 说明总体相关系数与 0 之间的差别具有统计学意义; 由于 $\tau = 0.56680 > 0$, 说明两种检测方法检测的结果之间呈正相关关系, 即两种方法检测的结果相差不会太大。

16.5 简单线性回归分析

16.5.1 对例 16-4 资料的分析

1. SAS 程序中重要内容的说明

程序名为 SASTJFX16_4. SAS。

```
data SASTJFX16_4;input x y @@ ;
cards;
0      80
0.8    90
... ..
4.8    135
;

ods html;
proc gplot;
plot y*x;
run;proc corr;var x y;
run;proc reg;model y=x;run;
ods html close;
```

数据步，输入两变量 x、y，依次录入数据。过程步，“proc reg; model y = x;run;”语句是进行简单的线性回归分析。

2. 主要分析结果及解释

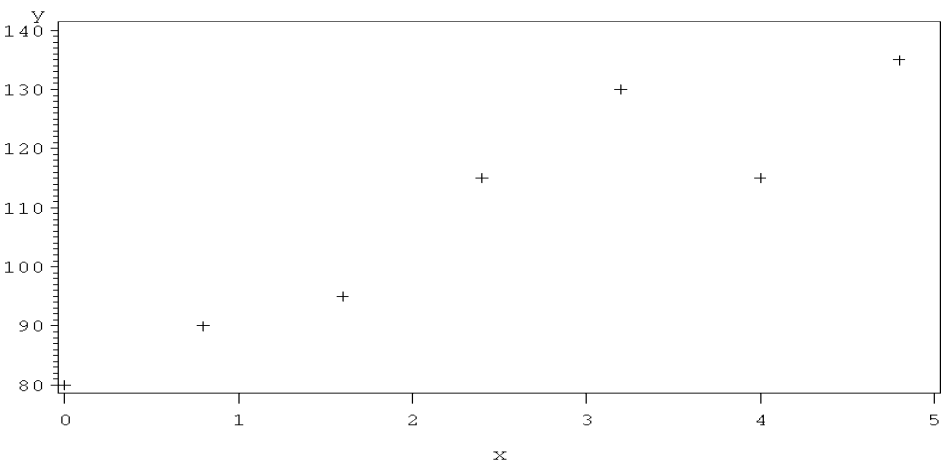


图 16-2 变量 y 与 x 的散点图

由图 16-2 的散点图可见，7 个点在一条不太宽的长带内随机地分布着，且不存在明显的曲线趋势，因此可以对该资料进行直线相关和回归分析。

| CORR 过程 | | |
|--------------------------|---------|---------|
| Pearson 相关系数，N = 7 | | |
| 当 H0:Rho = 0 时，Prob > r | | |
| | x | y |
| x | 1.00000 | 0.92912 |
| | | 0.0025 |
| y | 0.92912 | 1.00000 |
| | <.0025 | |

相关分析的结果： $r=0.92912$ ， $P=0.0025 < 0.05$ ，说明变量 x 与 y 之间存在相关性。

The REG Procedure

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|-----------------|----|----------------|-------------|---------|--------|
| Model | 1 | 2232.14286 | 2232.14286 | 31.57 | 0.0025 |
| Error | 5 | 353.57143 | 70.71429 | | |
| Corrected Total | 6 | 2585.71429 | | | |

以上为方差分析结果： $F=31.57$ ， $P=0.0025 < 0.05$ ，说明该直线回归方程在总体上具有统计学意义。

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
|-----------|----|--------------------|----------------|---------|---------|
| Intercept | 1 | 81.78571 | 5.72989 | 14.27 | <.0001 |
| x | 1 | 11.16071 | 1.98648 | 5.62 | 0.0025 |

由参数检验的结果得出：截距项 (Intercept) 与 0 之间的差别有统计学意义 ($t=14.27$ 、 $P < 0.0001$)，斜率与 0 之间的差别有统计学意义 ($t=5.62$ 、 $P=0.0025 < 0.05$)，因此建立直线回归方程为 $\hat{Y} = 81.78571 + 11.16071X$ ， X 为土壤 NaCl 含量， Y 为植物单位叶面积干物重量。由于 $r^2 = 0.8633 > 0.5$ ，说明此直线回归方程具有较高的实用价值。

16.5.2 对例 16-5 资料的分析

1. SAS 程序中重要内容的说明

程序名为 SASTJFX16_4A.SAS。

```
data SASTJFX16_4A;
input x y ;
cards;
40 60
41 61
... ..
62 100
;
run;
```

```
ods html;proc gplot;
plot y*x;run;proc corr;var x y;
run;proc reg;model y=x;run;
ods html close;
```

数据步，输入两变量 x 、 y ，依次录入数据。过程步，分别进行相关分析和回归分析。Model 语句指定回归模型的形式。

2. 主要分析结果及解释

图 16-3 的散点图表明 10 个点在一条不太宽的长带内随机分布，不存在明显的曲线趋势，可以考虑进行直线相关和回归分析。

CORR 过程

Pearson 相关系数， $N=10$

当 $H_0: \rho = 0$ 时， $\text{Prob} > |r|$

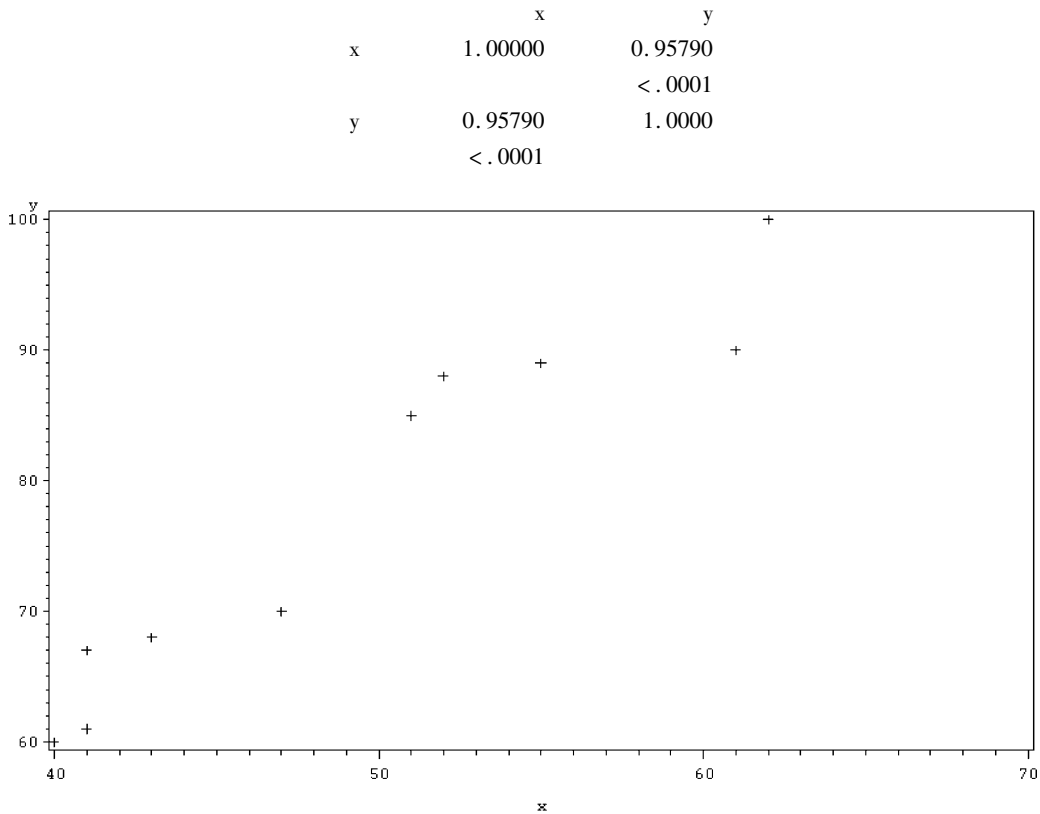


图 16-3 变量 y 与 x 的散点图

相关分析结果： $r=0.95790$ ， $P<0.0001$ ，变量 Y 与变量 X 之间有相关性。

| The REG Procedure | | | | | |
|----------------------|----|----------------|-------------|---------|--------|
| Analysis of Variance | | | | | |
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 1 | 1647.60983 | 1647.60983 | 89.07 | <.0001 |
| Error | 8 | 147.99017 | 18.49877 | | |
| Corrected Total | 9 | 1795.60000 | | | |

直线回归方程是否具有统计学意义的方差分析结果得到 $F=89.07$ ， $P<0.0001$ ，说明该直线回归方程在整体上具有统计学意义。

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | -3.21652 | 8.69163 | -0.37 | 0.7209 |
| x | 1 | 1.64334 | 0.17413 | 9.44 | <.0001 |

参数检验的结果可以得出：截距项 (Intercept) 与 0 之间差别无统计学意义， $P=0.7209>0.05$ ，说明该直线回归方程通过原点，因此需将截距项去掉重新拟合回归方程。

3. 回归方程中去掉截距项的 SAS 程序中重要内容的说明

程序名为 SASTJFX16_4B.SAS。

```

data SASTJFX16_4B;
input x y;
cards;
40 60
41 61
... ...
62 100
;
run;

ods html;
proc reg;
model y=x/NOINT R P;
run;
ods html close;

```

在程序 SASTJFX16_4B 中的“model y = x/NOINT R P;”语句中,“NOINT”的含义是去掉截距项拟合回归方程,“R”要求系统进行残差分析,“P”计算各自变量所对应的应变变量预测值。

4. SAS 程序修改后的主要分析结果及解释

The REG Procedure

Note: No intercept in model. R-Square is redefined.

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|-------------------|----|----------------|-------------|---------|--------|
| Model | 1 | 62173 | 62173 | 3717.43 | <.0001 |
| Error | 9 | 150.52362 | 16.72485 | | |
| Uncorrected Total | 10 | 62324 | | | |

以上是不包含截距项的直线回归方程统计学检验的结果,方差分析得 $F = 3717.43$, $P < 0.0001$,说明拟合的直线回归方程在整体上来说具有统计学意义

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
|----------|----|--------------------|----------------|---------|---------|
| x | 1 | 1.57969 | 0.02591 | 60.97 | <.0001 |

以上给出了不包含截距项的直线回归方程的参数估计值并对其假设检验的结果。所求得的回归方程为: $\hat{Y} = 1.57969X$ (注: Y 为存活百分数, X 为处理株第一叶平均高度占对照株高度的百分数)。

Output Statistics

| Obs | Dependent Variable | Predicted Value | Std Error Mean Predict | Residual | Std Error Residual | Student Residual | -2 -1 0 1 2 | Cook's D |
|-----|--------------------|-----------------|------------------------|----------|--------------------|------------------|-------------|----------|
| 1 | 60.0000 | 63.1876 | 1.0364 | -3.1876 | 3.956 | -0.806 | * | 0.045 |
| 2 | 61.0000 | 64.7673 | 1.0623 | -3.7673 | 3.949 | -0.954 | * | 0.066 |
| 3 | 67.0000 | 64.7673 | 1.0623 | 2.2327 | 3.949 | 0.565 | * | 0.023 |
| 4 | 68.0000 | 67.9267 | 1.1141 | 0.0733 | 3.935 | 0.0186 | | 0.000 |
| 5 | 70.0000 | 74.2455 | 1.2177 | -4.2455 | 3.904 | -1.087 | * * | 0.115 |
| 6 | 85.0000 | 80.5642 | 1.3214 | 4.4358 | 3.870 | 1.146 | * * | 0.153 |
| 7 | 88.0000 | 82.1439 | 1.3473 | 5.8561 | 3.861 | 1.517 | * * * | 0.280 |
| 8 | 89.0000 | 86.8830 | 1.4250 | 2.1170 | 3.833 | 0.552 | * | 0.042 |
| 9 | 90.0000 | 96.3611 | 1.5804 | -6.3611 | 3.772 | -1.686 | * * * | 0.499 |
| 10 | 100.0000 | 97.9408 | 1.6064 | 2.0592 | 3.761 | 0.548 | * | 0.055 |

以上是与各观测点对应的“应变量 Y 的观测值(第 2 列)”、“应变量 Y 的预测值(第 3 列)”、“应变量 Y 的预测值的标准误(第 4 列)”、“应变量 Y 的观测值与预测值之差,被称为残差(第 5 列)”、“残差的标准误(第 6 列)”、“学生化残差,即残差除以其标准误(第 7 列)”、“学生化残差图,学生化残差绝对值大于 2 的点将用 4 个‘*’号表示出来(若有这样的点),将被视为可疑的异常点(第 8 列)”、“CookD 距离统计量(第 8 列),此值越大表明该点越有可能是可疑的异常点”。

专业结论: 结合专业知识和散点图,可以对该资料进行直线相关与回归分析,所得直线相关系数 $r=0.95790$, $P<0.0001$,说明处理株第一叶平均高度占对照株高度的百分数(X)和存活百分数(Y)之间的直线相关关系有统计学意义。由于决定系数 $r^2=0.9176>0.5$,说明若根据此资料建立用 X 推测 Y 的直线回归方程有一定的实用价值,自变量 X 对因变量 Y 的贡献率约为 91.76%,即 Y 的变化中有 91.76% 的量可由 X 的变化来解释。求得由处理株第一叶平均高度占对照株高度的百分数(X)推算存活百分数(Y)的直线回归方程为: $\hat{Y}=1.57969X$ 。去掉截距项后的决定系数增至 0.9976,此直线回归方程具有较大的实用价值。

16.6 常用于估计 LD50 的加权线性回归分析

16.6.1 SAS 程序中重要内容的说明

程序名为 SASTJFX16_5A.SAS。

```
data SASTJFX16_5A;
input a dose r n;
cards;
1 0.3 0 8
1 0.4 2 8
.....
3 1.0 9 10
;
run;
```

```
ods html;
PROC PROBIT LOG10;
model r/n = dose/
LACKFIT INVERSECL;
by a;
run;
ods html close;
```

程序 SASTJFX16_5A 中“proc PROBIT LOG10; model r/n = dose/ LACKFIT INVERSECL;”,该语句调用 PROBIT 过程,选择项 LOG10 是对剂量取常用对数,MODEL 语句等号右边的 LACKFIT 要求对失拟进行检验,INVERSECL 要求求出用原始剂量所表达的各种反应效应量。

16.6.2 主要分析结果及解释

| Probit Procedure | | | |
|-----------------------|--------|----|------------|
| a = 1 | | | |
| Goodness-of-Fit Tests | | | |
| Statistic | Value | DF | Pr > ChiSq |
| Pearson Chi-Square | 4.7528 | 4 | 0.3136 |
| L. R. Chi-Square | 4.7009 | 4 | 0.3194 |

用 2 种方法对第 1 批资料进行失拟检验的结果, Pearson Chi-Square 值为 4.7528, P 值为 0.3136, L. R. Chi-Square 值为 4.7009, P 值为 0.3194,表明用加权的直线回归方程描述此资料是合适的。

| Analysis of Parameter Estimates | | | | | | | |
|---------------------------------|----|----------|----------------|-----------------------|---------|------------|------------|
| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
| Intercept | 1 | 2.3740 | 0.5274 | 1.3403 | 3.4077 | 20.26 | <.0001 |
| Log10(dose) | 1 | 9.3812 | 2.0501 | 5.3631 | 13.3993 | 20.94 | <.0001 |

求得第 1 批资料的加权直线回归方程为: $\hat{Y} = 2.3740 + 9.3812\lg(\text{dose})$ 。此式中的 \hat{Y} 为概率单位的预测值, 对截距和斜率的检验结果均为 $P < 0.0001$ 。

Probit Model in Terms of
Tolerance Distribution
MU SIGMA
-0.2530604 0.10659599

$\mu = -0.25306$ 是刺激(此处指对数半数致死剂量)的均数, $\sigma = 0.106596$ 是刺激的尺度参数。所求得的直线回归方程中截距 a 、斜率 b 与 μ 、 σ 之间的关系如下: $a = -\mu/\sigma$ 、 $b = 1/\sigma$ 。

| Probit Procedure | | | |
|-----------------------|--------|----|------------|
| a = 2 | | | |
| Goodness-of-Fit Tests | | | |
| Statistic | Value | DF | Pr > ChiSq |
| Pearson Chi-Square | 0.4964 | 3 | 0.9197 |
| L. R. Chi-Square | 0.5107 | 3 | 0.9165 |

用 2 种方法对第 2 批资料进行失拟检验的结果, Pearson Chi-Square 值为 0.4964, P 值为 0.9197, L. R. Chi-Square 值为 0.5107, P 值为 0.9165, 表明用加权的直线回归方程描述此资料是合适的。

| Analysis of Parameter Estimates | | | | | | | |
|---------------------------------|----|----------|----------------|-----------------------|---------|------------|------------|
| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
| Intercept | 1 | -2.9650 | 0.8107 | -4.5541 | -1.3760 | 13.38 | 0.0003 |
| Log10(dose) | 1 | 8.4315 | 2.2395 | 4.0421 | 26.8208 | 14.17 | 0.0002 |

求得第 2 批资料的加权直线回归方程为: $\hat{Y} = -2.9650 + 8.4315\lg(\text{dose})$ 。对截距的检验结果为 $P = 0.0003$, 对斜率的检验结果为 $P = 0.0002$ 。

Probit Model in Terms of
Tolerance Distribution
MU SIGMA
0.35166428 0.11860329

$\mu = 0.35166$ 是刺激(此处指对数半数致死剂量)的均数, $\sigma = 0.118603$ 是刺激的尺度参数。

| Probit Procedure | | | |
|-----------------------|--------|----|------------|
| a = 3 | | | |
| Goodness-of-Fit Tests | | | |
| Statistic | Value | DF | Pr > ChiSq |
| Pearson Chi-Square | 0.4263 | 2 | 0.8080 |
| L. R. Chi-Square | 0.5103 | 2 | 0.7748 |

用2种方法对第3批资料进行失拟检验的结果, Pearson Chi-Square 值为0.4263, P 值为0.8080, L. R. Chi-Square 值为0.5103, P 值为0.7748, 表明用加权的直线回归方程描述此资料是合适的。

| Analysis of Parameter Estimates | | | | | | | |
|---------------------------------|----|----------|----------------|-----------------------|---------|------------|------------|
| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
| Intercept | 1 | 1.1879 | 0.4259 | 0.3532 | 2.0227 | 7.78 | 0.0053 |
| Log10(dose) | 1 | 22.7999 | 6.2447 | 10.5606 | 35.0393 | 13.33 | 0.0003 |

求得第3批资料的加权直线回归方程为: $\hat{Y} = 1.1879 + 22.7999 \lg(\text{dose})$ 。对截距的检验结果为 $P=0.0053$, 对斜率的检验结果为 $P=0.0003$ 。

| Probit Model in Terms of Tolerance Distribution | |
|--|------------|
| MU | SIGMA |
| -0.0521031 | 0.04385976 |

$\mu = -0.0521031$ 是刺激(此处指对数半数致死剂量)的均数, $\sigma = 0.04385976$ 是刺激的尺度参数。

上述程序只能分别求出3种毒物所对应的加权直线回归方程及各种效应量(含 LD_{50}), 关于 LD_{50} 之间的比较、斜率之间的比较, 需用下面的程序来实现。具体 SAS 程序(程序名为 SAS-TJFX16_5B. SAS):

16.6.3 用于比较 LD_{50} 和斜率的 SAS 程序中重要内容的说明

```
ods html;
DATA SASTJFX16_5B;
  xld50a = -0.25306; xld50b = 0.35166; xld50c = -0.052103;
  ua = -0.21413;   la = -0.29682;
  ub = 0.41571;   lb = 0.29088;
  uc = -0.024937; lc = -0.077421;
  b1 = 9.38121595; b2 = 8.43146945; b3 = 22.799945;
  sb1 = 2.050085;  sb2 = 2.239507;  sb3 = 6.244696;
  q = 2 * 1.96;
  sld50a = (ua - la) / q; sld50b = (ub - lb) / q; sld50c = (uc - lc) / q;
  uab = abs(xld50a - xld50b) / sqrt(sld50a**2 + sld50b**2);
  uac = abs(xld50a - xld50c) / sqrt(sld50a**2 + sld50c**2);
  pld50ab = (1 - probnorm(uab)) * 2;
  pld50ac = (1 - probnorm(uac)) * 2;
  tab = abs(b1 - b2) / sqrt(sb1**2 + sb2**2);
  tac = abs(b1 - b3) / sqrt(sb1**2 + sb3**2);
  p_b_ab = (1 - probnorm(tab)) * 2;
  p_b_ac = (1 - probnorm(tac)) * 2;
  FILE PRINT;
  PUT #2 @ 10 'xld50a'   @ 25 'xld50b'   @ 40 'xld50c'
      #3 @ 10 xld50a 10.6 @ 25 xld50b 10.6 @ 40 xld50c 10.6
      #4 @ 10 'sld50a'   @ 25 'sld50b'   @ 40 'sld50c'
      #5 @ 10 sld50a 10.6 @ 25 sld50b 10.6 @ 40 sld50c 10.6
```



```

#6 @ 10 'uab' @ 25 'uac' @ 40 'pld50ab' @ 55 'pld50ac'
#7 @ 10 uab 5.3 @ 25 uac 5.3 @ 40 pld50ab 6.4 @ 55 pld50ac 6.4
#9 @ 10 'b1' @ 25 'b2' @ 40 'b3'
#10 @ 10 b1 10.6 @ 25 b2 10.6 @ 40 b3 10.6
#11 @ 10 'sb1' @ 25 'sb2' @ 40 'sb3'
#12 @ 10 sb1 10.6 @ 25 sb2 10.6 @ 40 sb3 10.6
#13 @ 10 'tab' @ 25 'tac' @ 40 'p_b_ab' @ 55 'p_b_ac'
#14 @ 10 tab 5.3 @ 25 tac 5.3 @ 40 p_b_ab 6.4 @ 55 p_b_ac 6.4;
RUN;
ods html close;

```

此程序的目的是对多个 LD_{50} 进行两两比较、对多个直线斜率进行两两比较。由于 PROBIT 过程不能直接实现此目的,需用 SAS 语言和统计公式来间接实现。首先,需将例 16-5 中输出的有关数据作为已知条件,赋给相应的变量。各变量含义如下:

xld50a、xld50b、xld50c 分别是 a, b, c 3 批数据的对数半数致死剂量;

ua、la 分别是第 1 组资料对数半数致死剂量的 95% 置信限的上限与下限,同理,知 ub, lb, uc, lc 的含义;

b1 ~ b3 分别是 3 条回归直线的样本斜率、sb1 ~ sb3 分别是 b1 ~ b3 的标准误差。

16.6.4 两两比较的主要分析结果及解释

| | | | |
|-----------|----------|-----------|---------|
| xld50a | xld50b | xld50c | |
| -0.253060 | 0.351660 | -0.052103 | |
| sld50a | sld50b | sld50c | |
| 0.021094 | 0.031844 | 0.013389 | |
| uab | uac | pld50ab | pld50ac |
| 15.83 | 8.043 | 0.0000 | 0.0000 |
| b1 | b2 | b3 | |
| 9.381216 | 8.431469 | 22.799945 | |
| sb1 | sb2 | sb3 | |
| 2.050085 | 2.239507 | 6.244696 | |
| tab | tac | p_b_ab | p_b_ac |
| 0.313 | 2.042 | 0.7544 | 0.0412 |

xld50a ~ xld50c、sld50a ~ sld50c 分别是 3 批数据对数半数致死剂量及其标准误;其后是第 1 组与第 2 组(标志分别为 a, b)、第 1 与第 3 组(标志分别为 a, c)之间对数半数致死剂量比较的 U 检验结果:前者为 $U = 15.83$, $P < 0.0001$;后者为 $U = 8.043$, $P < 0.0001$ 。

b1 ~ b3、sb1 ~ sb3 分别是 3 条回归直线的斜率及其标准误;其后是 b1 与 b2 比较、b1 与 b3 比较的 t 检验的结果:前者为 $t = 0.313$ 、 $P = 0.7544$, 差别无统计学意义;后者为 $t = 2.042$ 、 $P = 0.0412$, 差别具有统计学意义。

专业结论:与对照毒物($ld_{50} = 0.55839$)相比,第 1 种未知毒物的毒力($ld_{50} = 2.24732$)低于对照毒物;第 2 种未知毒物的毒力($ld_{50} = 0.886946$)也低于对照毒物。与对照毒物的回归斜率($b_1 \approx 9.38$)相比,第 1 种未知毒物的斜率($b_2 \approx 8.43$)与对照毒物的斜率差别无统计学意义;第 2 种未知毒物的斜率($b_3 \approx 22.80$)大于对照毒物的斜率。

16.7 本章小结

研究两个变量之间的相互关系和依赖关系，通常需要运用简单相关与回归分析(曲线相关与回归分析超出了本章范围)。在运用的过程中，应把握以下几点：第一，要特别注意同质性，即对于特定的研究目的，资料应来自性质相同的总体；第二，所研究的变量在专业上是有联系的，这种联系是以专业知识为依据的；第三，应绘制并分析反映两定量变量变化趋势的散点图，对值得进一步分析的资料再继续完成相应的计算和假设检验；第四，相关与回归分析的结果是否有应用价值，应结合决定系数 r^2 和回归预测结果来综合评价。本章详细介绍了 SAS 软件 CORR、REG 两个过程进行简单相关与回归分析的调用方法，给出了一些实例以及相关的 SAS 程序、分析方法、运行结果和结果解释。对于半数效量估计的加权回归分析结合实例调用 PROBIT 过程进行了分析和结果解释。

(郭晋 陶丽新)

第 17 章 两变量可直线化曲线回归分析

在科学研究中,研究者经常希望考察两个在专业上有联系的变量或指标之间的关系。此时,可采用回归分析。当两个变量之间存在直线关系且满足简单直线回归应用的前提条件时,可采用简单直线回归的方法进行分析。但很多时候,两个变量之间并不呈直线关系,而是呈非线性关系或称曲线关系,这时选择恰当的曲线比拟合直线更为合适。如细菌繁殖与培养时间的关系,疾病发病率与时间的关系,毒物剂量与反应之间的关系等通常不是直线关系。本章介绍如何借助 SAS 软件实现两变量可直线化曲线回归分析。

17.1 问题、数据及统计分析方法的选择

17.1.1 问题与数据

【例 17-1】 上海医科大学微生物教研室以已知浓度的免疫球蛋白 A(IgA, μg) 做火箭电泳,测得火箭的高度(mm)如下,试拟合其曲线回归方程。

| | | | | | | | | |
|----------------------------|-----|------|------|------|------|------|------|------|
| IgA 浓度(x, μg): | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 | 1.2 | 1.4 | 1.6 |
| 火箭高度(y, mm): | 7.6 | 12.3 | 15.7 | 18.2 | 18.7 | 21.4 | 22.6 | 23.8 |

【例 17-2】 某研究者收集了一只红铃虫的产卵数 y 和温度 x 之间的 7 组观测数据,资料如下。试拟合其曲线回归方程。

| | | | | | | | |
|--------------------------------|----|----|----|----|----|-----|-----|
| 温度 x ($^{\circ}\text{C}$): | 21 | 23 | 25 | 27 | 29 | 32 | 35 |
| 产卵数 y (个): | 7 | 11 | 21 | 24 | 66 | 115 | 325 |

【例 17-3】 某研究组调查某县 1980—1996 年月累计麻疹的发病情况,观察指标为发病率(1/10 万),1~12 月(x) 累计的发病率(y) 依次为: 0.279、0.920、2.078、4.393、8.144、12.174、17.306、18.435、18.728、18.965、19.230、19.328。试拟合其曲线回归方程。

17.1.2 对数据结构的分析

从形式上看,前述三个实例所代表呈现的数据结构是完全相同的,其数据结构属于单组设计定量资料,结合实际问题,可认为它们都含有一个自变量 x 和一个因变量 y ; 但三组数据的表现是不尽相同的,这一点需要通过绘制出反映两个变量同时变化趋势的散点图(参见图 17-1~图 17-3)才能有一个直观印象。

17.1.3 分析目的与统计分析方法的选择

对例 17-1 资料而言,为考察响应变量随自变量变化而变化的趋势,可以使用回归分析。要对一个资料拟合回归方程,首先应绘制散点图,以了解各散点的变化趋势,在对散点图进行观察后,发现该资料的散点图呈现明显的曲线形状,因而考虑拟合曲线回归方程。进一步观

察，散点的变化趋势近似于对数曲线、双曲线或幂函数曲线，故采用对数变换法或倒数变换法使其直线化。曲线直线化变量变换的方法如表 17-1 所示。

表 17-1 常见曲线直线化变量变换的方法

| 曲线类型 | 曲线方程 | 变换方法 | 直线化结果 |
|-------------|---------------------------------|--------------------------------------|--------------------|
| 对数曲线 | $y = a + b \lg(x)$ | $x' = \lg(x)$ | $y = a + bx'$ |
| | $y = a + b \lg(x + k)$ | $x' = \lg(x + k)$ | |
| 指数曲线 | $y = ae^{bx}$ | $y' = \ln y$ | $y' = \ln a + bx$ |
| | $y = a10^{bx}$ | $y' = \lg y$ | $y' = \lg a + bx$ |
| 幂函数曲线 | $y = ax^b$ | $x' = \ln x, y' = \ln y$ | $y' = \ln a + bx'$ |
| 双曲线 | $\frac{1}{y} = a + \frac{b}{x}$ | $x' = \frac{1}{x}, y' = \frac{1}{y}$ | $y' = a + bx'$ |
| logistic 曲线 | $y = L + \frac{K}{1 + ae^{bx}}$ | $y' = \ln \frac{K - (y - L)}{y - L}$ | $y' = bx$ |

对例 17-2 资料而言，绘制散点图可发现，散点的变化趋势近似于指数曲线，故采用对因变量取对数变换的方法使其直线化。曲线直线化变量变换的方法如表 17-1 所示。

对例 17-3 资料而言，绘制散点图可发现，散点的变化趋势近似于“S”形曲线，故采用 logit 变换使其直线化。曲线直线化变量变换的方法如表 17-1 所示。

17.2 对数函数、幂函数和双曲线函数曲线回归分析

17.2.1 SAS 程序中重要内容的说明

例 17-1 资料拟合曲线方法程序，程序名称为 SASTJFX17_1A.SAS。

```
DATA SASTJFX17_1A;          /* 1 */
  INPUT x y @@ ;
  x1 = log10(x);
  x2 = 1/x;
  x3 = log(x);
  y1 = 1/y;
  y2 = log(y);
  CARDS;
  0.2  7.6  0.4  12.3  0.6  15.7
  0.8  18.2  1.0  18.7  1.2  21.4
  1.4  22.6  1.6  23.8
  ;
RUN;

ods html;
PROC GPLOT;                  /* 2 */
  PLOT y*x/haxis=0 to 1.6 by 0.4
       vaxis=0 to 25 by 5;
  symbol value = dot;
RUN;
PROC REG;                    /* 3 */
  model y = x1;
  model y1 = x2;
  model y2 = x3;
RUN;
ods html close;
```

程序第 1 步建立数据集，x 代表 IgA 浓度，y 代表火箭高度，并对两变量进行各种变换，以便进行曲线直线化分析。第 2 步绘制散点图，以判断散点趋势。第 3 步进行各种回归分析，包括对数曲线、双曲线和幂函数曲线的直线化分析。

例 17-1 资料曲线拟合结果情况分析，程序名称为 SASTJFX17_1B.SAS。

```
%let n=8;                   /* 1 */
data;                        /* 2 */
  input x y@@ ;
  y1 = 19.74512 + 17.90734
       * log10(x);

insert into test1 select SS from
a where Source = 'Corrected
Total';
insert into test2 select
sum(ssrs) from;
```

```

residual=y1-y;
ssrs=residual**2;
cards;
0.2  7.6  0.4  12.3  0.6  15.7
0.8  18.2  1.0  18.7  1.2  21.4
1.4  22.6  1.6  23.8
;
run;

ods html;
proc print;                                /* 3 */
run;
ods html close;

proc reg;                                /* 4 */
model y=x;
ods output ANOVA=a;
run;

proc sql;                                /* 5 */
create table Test1(total num);
create table Test2(sum_resi num);
run;

run;
quit;

data test;                                /* 6 */
merge test1 test2;
s=sqrt(sum_resi/(&n-2));
r_square=1-sum_resi/total;
run;

ods html;
proc print noobs;                          /* 7 */
run;
ods html close;

```

第 1 步设立宏变量 n ，其含义为样本个数。第 2 步创建数据集，并计算曲线 $y1 = 19.74512 + 17.90734 \times \log_{10}(x)$ 拟合资料的残差 residual 及残差平方和 ssrs 。第 3 步将第 2 步所得到的数据集输出到 output 窗口。第 4 步对 x 与 y 进行简单直线回归分析，目的在于得到变量 y 的离均差平方和。第 5 步通过 sql 过程创立两个表格 Test1 和 Test2 ，两表格各包含一个变量。将由数据集 a 中的变量 y 的离均差平方和插入表格 Test1 ，将由数据集 sastjfx8_2 计算而得的残差平方和插入表格 Test2 。第 6 步合并表格 Test1 和 Test2 ，并计算新变量 s 和 r_square ，分别代表剩余标准差和相关指数。第 7 步输出数据集 test 至 output 窗口(注意：这里所讲的表格实际上就是数据集)。

17.2.2 主要分析结果及解释

图 17-1 是根据原始资料绘制的 IgA 浓度和火箭高度之间数据对的散点图，可发现其趋势与对数曲线、双曲线或幂函数曲线很相似，故对两变量进行相应变换。

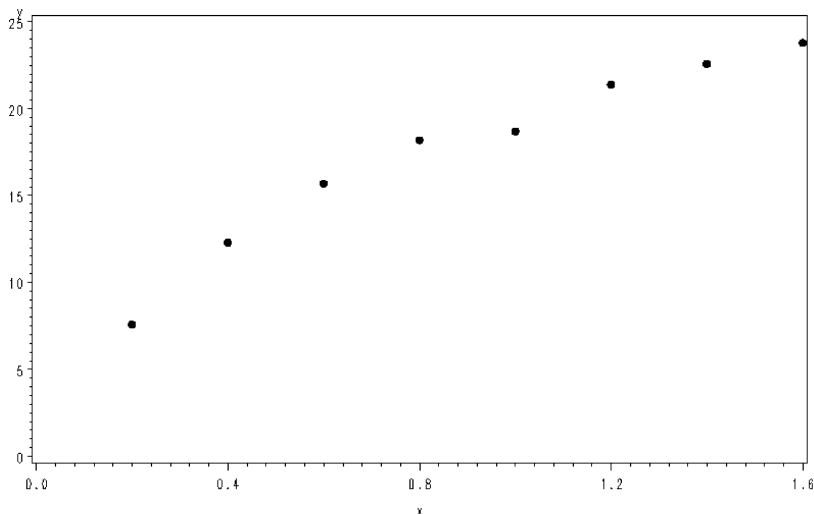


图 17-1 例 17-1 资料的散点图

| The REG Procedure | | | | | |
|-----------------------|----|----------------|-------------|---------|--------|
| Model: MODEL1 | | | | | |
| Dependent Variable: y | | | | | |
| Analysis of Variance | | | | | |
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 1 | 209.47260 | 209.47260 | 763.50 | <.0001 |
| Error | 6 | 1.64615 | 0.27436 | | |
| Corrected Total | 7 | 211.11875 | | | |
| Root MSE | | 0.52379 | R-Square | 0.9922 | |
| Dependent Mean | | 17.53750 | Adj R-Sq | 0.9909 | |
| Coeff Var | | 2.98670 | | | |

这是响应变量火箭高度(y)和 IgA 浓度(x)经对数变换后产生的新变量(x1)之间进行简单直线回归的结果。由第一部分的方差分析结果可知, $F = 763.50$, $P < .0001$, 说明 y 与 x1 之间拟合的直线回归方程有统计学意义。

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 19.74512 | 0.20169 | 97.90 | <.0001 |
| x1 | 1 | 17.90734 | 0.64808 | 27.63 | <.0001 |

这是拟合的直线回归方程中的截距项和斜率的估计值, 以及它们与零的差别进行假设检验的结果。可以看出, 截距和斜率与零的差别均有统计学差异。因此, 直线回归方程为: $\hat{y} = 19.74512 + 17.90734x_1$ 。因 $x_1 = \log_{10}(x)$, 故 $\hat{y} = 19.74512 + 17.90734 \log_{10}(x)$ 。

| Model: MODEL2 | | | | | |
|------------------------|----|----------------|-------------|---------|--------|
| Dependent Variable: y1 | | | | | |
| Analysis of Variance | | | | | |
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 1 | 0.00622 | 0.00622 | 3524.42 | <.0001 |
| Error | 6 | 0.00001058 | 0.00000176 | | |
| Corrected Total | 7 | 0.00623 | | | |
| Root MSE | | 0.00133 | R-Square | 0.9983 | |
| Dependent Mean | | 0.06475 | Adj R-Sq | 0.9980 | |
| Coeff Var | | 2.05103 | | | |

这是响应变量火箭高度(y)经倒数变换后产生的新变量(y1)和 IgA 浓度(x)经倒数变换后产生的新变量(x2)之间进行简单直线回归的结果。由第一部分的方差分析结果可知, $F = 3524.42$, $P < .0001$, 说明 y1 与 x2 之间拟合的直线回归方程有统计学意义。

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |

| | | | | | |
|-----------|---|---------|------------|-------|--------|
| Intercept | 1 | 0.03029 | 0.00074660 | 40.57 | <.0001 |
| x2 | 1 | 0.02029 | 0.00034173 | 59.37 | <.0001 |

这是拟合的直线回归方程中的截距项和斜率的估计值, 以及它们与零的差别进行假设检验的结果。可以看出, 截距和斜率与零的差别均有统计学差异。因此, 直线回归方程为: $\hat{y}_1 = 0.03029 + 0.02029x_2$ 。因 $x_2 = 1/x$, $y_1 = 1/y$, 故 $\hat{y} = x/(0.03029x + 0.02029)$ 。

Model: MODEL3

Dependent Variable: y2

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|-----------------|----|----------------|-------------|---------|--------|
| Model | 1 | 0.99750 | 0.99750 | 332.17 | <.0001 |
| Error | 6 | 0.01802 | 0.00300 | | |
| Corrected Total | 7 | 1.01552 | | | |
| Root MSE | | 0.05480 | R-Square | 0.9823 | |
| Dependent Mean | | 2.80905 | Adj R-Sq | 0.9793 | |
| Coeff Var | | 1.95081 | | | |

这是响应变量火箭高度(y)经对数变换后产生的新变量(y2)和 IgA 浓度(x)经对数变换后产生的新变量(x3)之间进行简单直线回归的结果。由第一部分的方差分析结果可知, $F = 332.17$, $P < .0001$, 说明 y_2 与 x_3 之间拟合的直线回归方程有统计学意义。

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
|-----------|----|--------------------|----------------|---------|---------|
| Intercept | 1 | 2.96139 | 0.02110 | 140.35 | <.0001 |
| x3 | 1 | 0.53667 | 0.02945 | 18.23 | <.0001 |

这是拟合的直线回归方程中的截距项和斜率的估计值, 以及它们与零的差别进行假设检验的结果。可以看出, 截距和斜率与零的差别均有统计学差异。因此, 直线回归方程为: $\hat{y}_2 = 2.96139 + 0.53667x_3$ 。因 $x_3 = \ln(x)$, $y_2 = \ln(y)$, 故 $\hat{y} = \exp(2.96139 + 0.53667\ln(x))$ 。

| Obs | x | y | y1 | residual | ssrs |
|-----|-----|------|---------|----------|---------|
| 1 | 0.2 | 7.6 | 7.2284 | -0.37157 | 0.13807 |
| 2 | 0.4 | 12.3 | 12.6191 | 0.31907 | 0.10181 |
| 3 | 0.6 | 15.7 | 15.7724 | 0.07240 | 0.00524 |
| 4 | 0.8 | 18.2 | 18.0097 | -0.19028 | 0.03621 |
| 5 | 1.0 | 18.7 | 19.7451 | 1.04512 | 1.09228 |
| 6 | 1.2 | 21.4 | 21.1630 | -0.23695 | 0.05615 |
| 7 | 1.4 | 22.6 | 22.3619 | -0.23812 | 0.05670 |
| 8 | 1.6 | 23.8 | 23.4004 | -0.39963 | 0.15971 |

这是拟合的曲线方程对结果变量的拟合情况, x 、 y 分别代表原始资料中的自变量与响应变量, residual 表示拟合的残差, 即估计值与真实值之差, ssrs 为残差平方。后同, 不再赘述。

| Obs | total | sum_resi | s | r_square |
|-----|---------|----------|---------|----------|
| 1 | 211.119 | 1.64615 | 0.52379 | 0.99220 |

这是衡量曲线拟合情况的几个统计量的值，total 为变量 y 的离均差平方和(计算相关指数用)，sum_resi 表示残差平方和，s 为剩余标准差，r_square 代表相关指数。从后三个统计量的值可以看出，资料拟合比较理想。

当然，这是对拟合的对数曲线分析的结果。若将 SAS 程序 SASTJFX13_2 中 DATA 步中的“y1 = 19.74512 + 17.90734 * log10(x);”分别替换成“y1 = x/(0.03029 * x + 0.02029);”及“y1 = exp(2.96139 + 0.53667 * log(x));”，就可得到双曲线函数及幂函数拟合的有关情况，依次为：

| | | | | |
|-----|---------|----------|---------|----------|
| Obs | total | sum_resi | s | r_square |
| 1 | 211.119 | 1.60966 | 0.51795 | 0.99238 |
| Obs | total | sum_resi | s | r_square |
| 1 | 211.119 | 4.50772 | 0.86677 | 0.97865 |

若想比较三个曲线回归方程拟合此资料是否有统计学差异，可进行曲线拟合优度的比较。三个曲线回归方程的残差平方和分别为 1.64615、1.60966 和 4.50772，故可先选出最大的残差平方和与最小残差平方和进行比较。这两个残差平方和对应的曲线回归方程中均有 2 个待估计参数，故其自由度均为 $8 - 2 = 6$ 。

$$F = 4.50772 / 1.60966 = 2.80$$

查方差齐性检验用的 F 临界值表可知， $F_{0.05/2,6,6} = 5.82$ ，因 $F = 2.80 < 5.82$ ，所以这两个曲线回归方程拟合此资料没有统计学差异，当然由于刚才比较的是三个回归曲线中最大的残差平方和与最小残差平方和，而这两个残差平方和对应的曲线回归方程拟合此资料没有统计学差异，这也就意味着 3 个曲线回归方程拟合此资料均无统计学差异。

其实，若想从 3 条曲线中选择 1 条较好的加以应用，不需要进行曲线拟合优度的比较，直接查看各曲线模型对应的 F 值即可。由于双曲线函数模型对应的 F 值 3524.42 最大，相比其他两个模型稍好一些，故此资料采用双曲线函数拟合效果较好。

17.3 指数函数曲线回归分析

17.3.1 SAS 程序中重要内容的说明

例 17-2 资料的程序名为 SASTJFX17_2A.SAS(程序解释参考上例说明)。

| | |
|---|---|
| <pre>DATA SASTJFX17_2A; INPUT x y @@ ; y1 = log10(y); CARDS; 21 7 23 11 25 21 27 24 29 66 32 115 35 325 ; RUN; ods html;</pre> | <pre>PROC GPLOT; PLOT y*x/haxis=21 to 35 by 2 vaxis=0 to 350 by 50; symbol value=dot; RUN; PROC REG; model y1=x; RUN; ods html close;</pre> |
|---|---|

例 17-2 资料曲线拟合情况分析，程序名称为 SASTJFX17_2B.SAS。


```
DATA SASTJFX17_2B;
  input x y @@ ;
  y1 =10** (-1.67168+0.11814* x);
  residual=y1 - y;
  ssrs=residual** 2;
  cards;
  21  7  23  11  25  21  27  24
  29  66  32  115  35  325
;
RUN;
ods html;
proc print;
run;
ods html close;
PROC REG;
  model y=x;
  ods output ANOVA=a;
RUN;
```

```
proc sql;
  create table Test1(total num);
  create table Test2(sum_resi num);
  insert into test1 select SS from
a where Source='Corrected Total';
  insert into test2 select sum
(ssrs) as sum_resi
from Sastjfx17_2B;
run;
data test;
  merge test1 test2;
  s=sqrt(sum_resi/(7-2));
  r_square=1-sum_resi/total;
run;
ods html;
proc print;
run;
ods html close;
```

17.3.2 主要分析结果及解释

图 17-2 是根据原始资料绘制的产卵数和温度之间数据对的散点图，可发现其趋势与指数曲线很相似，故对产卵数进行以 10 为底的对数变换。

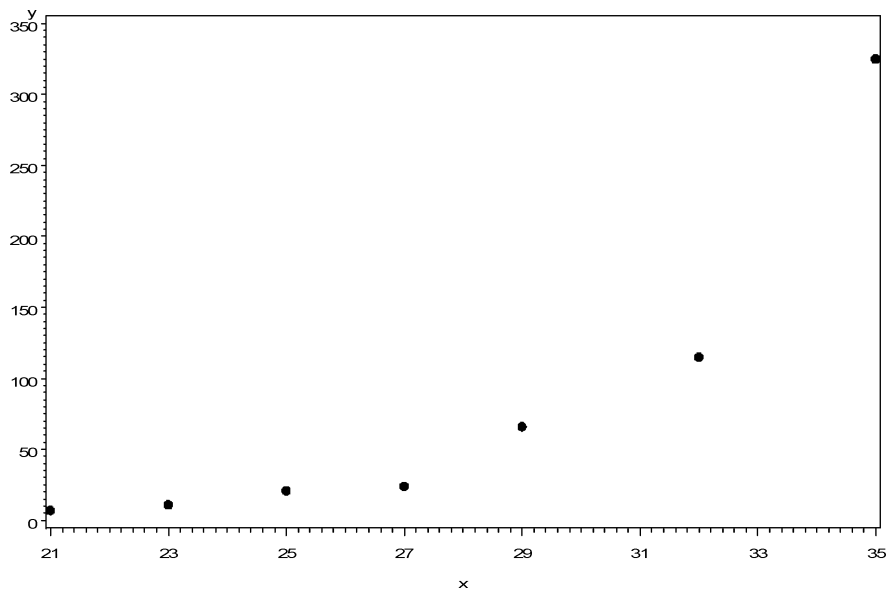


图 17-2 例 17-2 资料的散点图

Model: MODEL1

Dependent Variable: y1

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|--------|----|----------------|-------------|---------|---------|
| Model | 1 | 2.06164 | 2.06164 | 333.87 | < .0001 |

| | | | | | |
|-----------------|----------------|---------|----------|--------|--|
| Error | 5 | 0.03088 | 0.00618 | | |
| Corrected Total | 6 | 2.09251 | | | |
| | Root MSE | 0.07858 | R-Square | 0.9852 | |
| | Dependent Mean | 1.56872 | Adj R-Sq | 0.9823 | |
| | Coeff Var | 5.00927 | | | |

这是响应变量产卵数(y)经对数变换后产生的新变量(y1)和温度(x)之间进行简单直线回归的结果。由第一部分的方差分析结果可知, $F = 333.87$, $P < .0001$, 说明 y1 与 x 之间拟合的直线回归方程有统计学意义。

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | -1.67168 | 0.17981 | -9.30 | 0.0002 |
| x | 1 | 0.11814 | 0.00647 | 18.27 | <.0001 |

这是拟合的直线回归方程中的截距项和斜率的估计值, 以及它们与零的差别进行假设检验的结果。可以看出, 截距和斜率与零的差别均有统计学差异。因此, 直线回归方程为: $\hat{y}_1 = -1.67168 + 0.11814 x$ 。因 $y_1 = \log_{10}(y)$, 故 $\hat{y} = 10^{-1.67168 + 0.11814 x}$ 。

| Obs | x | y | y1 | residual | ssrs |
|-----|----------|----------|---------|----------|---------|
| 1 | 21 | 7 | 6.446 | -0.5544 | 0.31 |
| 2 | 23 | 11 | 11.106 | 0.1055 | 0.01 |
| 3 | 25 | 21 | 19.135 | -1.8654 | 3.48 |
| 4 | 27 | 24 | 32.969 | 8.9686 | 80.44 |
| 5 | 29 | 66 | 56.804 | -9.1959 | 84.56 |
| 6 | 32 | 115 | 128.469 | 13.4695 | 181.43 |
| 7 | 35 | 325 | 290.549 | -34.4506 | 1186.84 |
| Obs | total | sum_resi | s | r_square | |
| 1 | 78141.43 | 1537.07 | 17.5332 | 0.98033 | |

从残差平方和、剩余标准差及相关指数这三个统计量的值可以看出, 资料拟合比较理想。

17.4 logistic 函数曲线回归分析

17.4.1 SAS 程序中重要内容的说明

例 17-3 资料的程序名为 SASTJFX17_3A.SAS(程序解释参考上例说明)。

```
DATA SASTJFX17_3A;
  INPUT x y @@ ;
  y1=log((19.34-y)/y);
CARDS;
  1 0.279 2 0.920 3 2.078
  4 4.393 5 8.144 6 12.174
  7 17.306 8 18.435 9 18.728
  10 18.965 11 19.230 12 19.328
;
RUN;
```

```
ods html;
PROC GPLOT;
  PLOT y*x/haxis=0 to 12 by 3
  vaxis=0 to 20 by 2;
  symbol value=dot;
RUN;
PROC REG;
  model y1=x;
RUN;
ods html close;
```

例 17-3 资料曲线拟合情况分析，程序名为 SASTJFX17_3B.SAS。

```
DATA SASTJFX17_3B;
  input x y @@ ;
  y1 =19.34/ (exp (5.09934 -
    0.97293* x) +1);
  residual=y1 -y;
  ssrs=residual** 2;
  CARDS;
    1 0.279 2 0.920 3 2.078
    4 4.393 5 8.144 6 12.174
    7 17.306 8 18.435 9 18.728
    10 18.965 11 19.230 12 19.328
  ;
RUN;
ods html;
proc print;
run;
ods html close;
PROC REG;
  model y =x;
  ods output ANOVA =a;
RUN;
```

```
proc sql;
  create table Test1 (total num);
  create table Test2 (sum_resi num);
  insert into test1 select SS from a
  where Source = 'Corrected Total';
  insert into test2 select sum
    (ssrs)as sum_resi from Sastjfx17_3B;
run;
data test;
  merge test1 test2;
  s =sqrt (sum_resi/ (12 -2));
  r_square =1 -sum_resi/total;
run;
ods html;
proc print;
run;
ods html close;
```

17.4.2 主要分析结果及解释

图 17-3 是根据原始资料绘制的麻疹累计发病率和月份之间数据对的散点图，可发现其趋势近似 S 形，故对麻疹累计发病率进行 logit 变换。

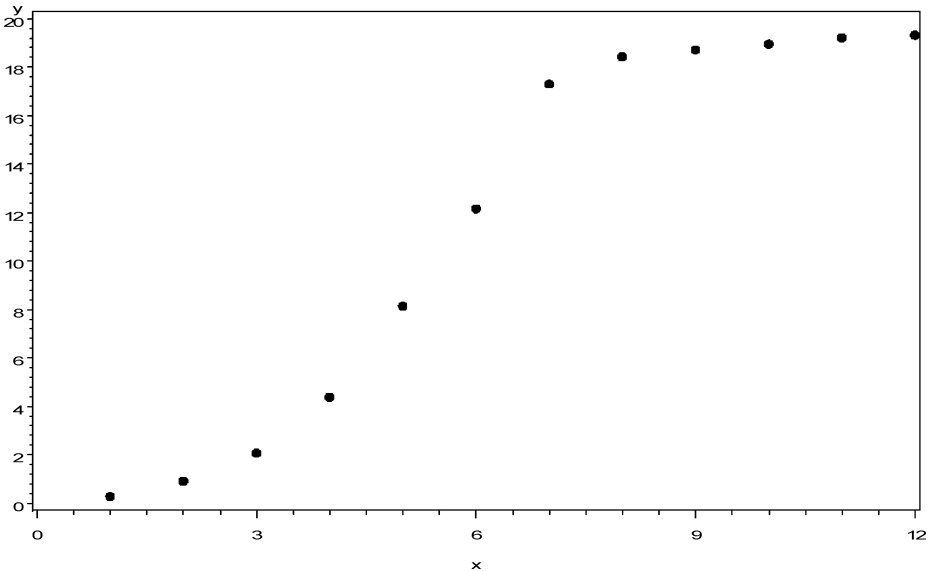


图 17-3 例 17-3 资料的散点图

Model: MODEL1

Dependent Variable: y1

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|--------|----|----------------|-------------|---------|--------|
|--------|----|----------------|-------------|---------|--------|

| | | | | | |
|-----------------|----|-----------|-----------|--------|--------|
| Model | 1 | 135.36387 | 135.36387 | 758.77 | <.0001 |
| Error | 10 | 1.78400 | 0.17840 | | |
| Corrected Total | 11 | 137.14787 | | | |
| Root MSE | | 0.42237 | R-Square | 0.9870 | |
| Dependent Mean | | -1.22473 | Adj R-Sq | 0.9857 | |
| Coeff Var | | -34.48698 | | | |

这是响应变量麻疹累计发病率(y)经 logit 变换后产生的新变量(y1)和月份(x)之间进行简单直线回归的结果。由第一部分的方差分析结果可知, $F = 758.77$, $P < .0001$, 说明 y1 与 x 之间拟合的直线回归方程有统计学意义。

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 5.09934 | 0.25995 | 19.62 | <.0001 |
| x | 1 | -0.97293 | 0.03532 | -27.55 | <.0001 |

这是拟合的直线回归方程中的截距项和斜率的估计值, 以及它们与零的差别进行假设检验的结果。可以看出, 截距和斜率与零的差别均有统计学差异。因此, 直线回归方程为: $\hat{y}1 = 5.09934 - 0.97293 x$ 。因 $y1 = \ln[(19.34 - y)/y]$, 故 $\hat{y} = 19.34 / [\exp(5.09934 - 0.97293x) + 1]$ 。

| Obs | x | y | y1 | residual | ssrs |
|-----|---------|----------|---------|----------|---------|
| 1 | 1 | 0.279 | 0.3072 | 0.02820 | 0.00080 |
| 2 | 2 | 0.920 | 0.7921 | -0.12794 | 0.01637 |
| 3 | 3 | 2.078 | 1.9632 | -0.11478 | 0.01317 |
| 4 | 4 | 4.393 | 4.4506 | 0.05757 | 0.00331 |
| 5 | 5 | 8.144 | 8.5405 | 0.39645 | 0.15718 |
| 6 | 6 | 12.174 | 13.0857 | 0.91165 | 0.83111 |
| 7 | 7 | 17.306 | 16.3807 | -0.92525 | 0.85609 |
| 8 | 8 | 18.435 | 18.1038 | -0.33118 | 0.10968 |
| 9 | 9 | 18.728 | 18.8534 | 0.12541 | 0.01573 |
| 10 | 10 | 18.965 | 19.1532 | 0.18816 | 0.03540 |
| 11 | 11 | 19.230 | 19.2690 | 0.03895 | 0.00152 |
| 12 | 12 | 19.328 | 19.3131 | -0.01492 | 0.00022 |
| Obs | total | sum_resi | s | r_square | |
| 1 | 699.326 | 2.04058 | 0.45173 | 0.99708 | |

从残差平方和、剩余标准差及相关指数这三个统计量的值可以看出, 资料拟合比较理想。

17.5 本章小结

本章介绍了可直线化曲线拟合的几种常用方法, 并给出了实例和用 SAS 解决具体问题的详细做法。值得注意的是, 适合进行曲线直线化分析的资料应满足特定的一些条件, 即资料属于单组设计二元定量资料, 资料所测自的受试者应具有同质性, 两个定量指标在专业上有一定联系, 绘制出反映两定量变量之间变化趋势的散点图呈现某种简单的曲线趋势(如 对数曲线、指数曲线、双曲线、幂函数曲线、S 形曲线或倒 S 形曲线), 此时采用本章介绍的方法进行曲线拟合效果一般比较理想。

(郭晋 陶丽新)

第 18 章 各种复杂曲线回归分析

在本书有关章节中介绍了两变量简单直线回归分析和两变量可直线化的曲线回归分析，这两种方法可进行一般的两变量间的回归分析。但生物医学研究中，变量间的回归关系往往并非这么简单，它们之间可能呈某种不易或不能被直线化的曲线关系；且曲线直线化是采用最小二乘法使变量转换后所得新变量离均差平方和最小，并不一定是原结果变量的离均差平方和最小，故所得模型的拟合精度仍有提高的空间。

本章主要介绍各种生物医学中可能用到的复杂曲线回归分析及其 SAS 实现，这包括多项式曲线回归分析、logistic 曲线回归分析、Gompertz 曲线回归分析和多项型指数曲线回归分析等。

18.1 多项式曲线回归分析

18.1.1 问题与数据

【例 18-1】 头围是反映婴幼儿脑和颅骨发育程度的重要指标之一。某研究者欲研究婴幼儿头围和身长之间的回归关系，于 1985 年调查了中国 9 个城市 2 岁以内婴幼儿体格的发育情况，不同月龄婴幼儿平均头围和平均身长见表 18-1。试拟合二者之间的曲线关系。

表 18-1 1985 年中国 9 个城市 2 岁以内婴幼儿的平均头围和平均身长

| 月龄(月) | 平均身长(cm) | 平均头围(cm) |
|-------|----------|----------|
| 初生 | 49.90 | 33.73 |
| 1~ | 56.60 | 37.45 |
| 2~ | 59.25 | 39.10 |
| ... | ... | ... |
| 24~ | 85.93 | 47.45 |

18.1.2 分析目的与统计分析方法的选择

绘制本资料的散点图(见图 18-1)可发现，散点的变化趋势近似于二项式曲线，故对本资料进行二项式曲线回归分析。

18.1.3 SAS 程序

SAS 程序名为 sastjfx18_1.SAS。

```
data sastjfx18_1;          /* 1 */
  input x y @@ ;
  z = x* *2;
  cards;
49.90 33.73 56.60 37.45
59.25 39.10 61.40 40.25
63.38 41.33 65.10 42.18
67.28 43.20 69.85 44.15
72.30 44.70 74.98 45.45
77.63 46.05 80.03 46.55
82.50 46.98 85.93 47.45
;
run;
```

```
ods html;
proc gplot;                /* 2 */
  plot y*x/haxis=40 to 90 by 10
  vaxis=30 to 50 by 5;
  symbol value = dot;
run;
proc reg;                  /* 3 */
  model y = x z;
run;
quit;
ods html close;
```

程序中，第 1 步建立数据集， x 代表平均身长， y 代表平均头围， z 为 y 的平方，在第 3 步对资料进行二项式曲线回归分析中将会用到。第 2 步绘制散点图。第 3 步对资料进行二项式曲线回归分析，以 z 代表 x 的平方项，从而将多项式曲线回归分析转换成多重线性回归分析。

18.1.4 主要分析结果及解释

图 18-1 是根据原始资料绘制的平均身长和平均头围之间数据对的散点图，可发现其趋势近似二次抛物线的一段，故对资料进行二项式曲线回归分析。

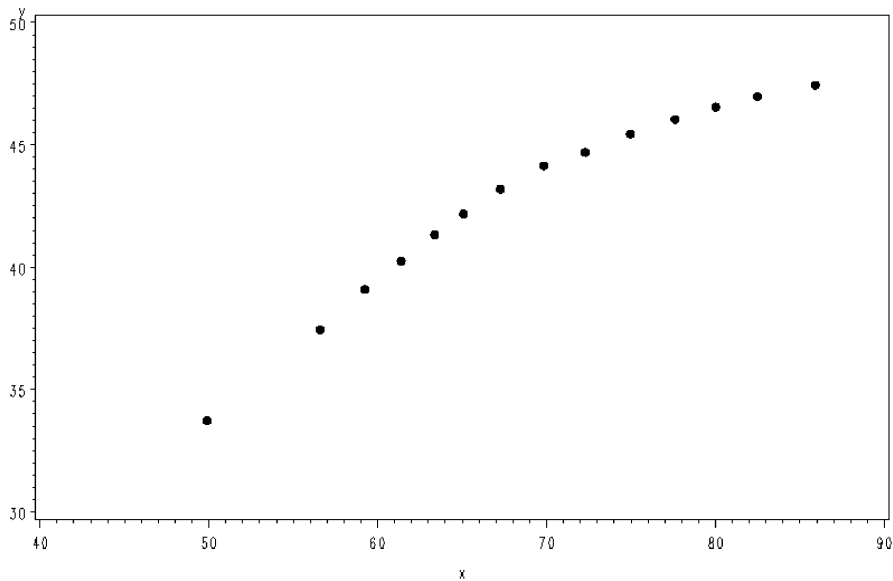


图 18-1 y 随 x 变化的散点图

| Analysis of Variance | | | | | |
|----------------------|----|--------------------|----------------|---------|---------|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 2 | 209.61061 | 104.80530 | 3589.84 | <.0001 |
| Error | 11 | 0.32114 | 0.02919 | | |
| Corrected Total | 13 | 209.93175 | | | |
| Root MSE | | 0.17087 | R-Square | 0.9985 | |
| Dependent Mean | | 42.75500 | Adj R-Sq | 0.9982 | |
| Coeff Var | | 0.39964 | | | |
| Parameter Estimates | | | | | |
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | -22.07048 | 1.99523 | -11.06 | <.0001 |
| x | 1 | 1.53801 | 0.05883 | 26.14 | <.0001 |
| z | 1 | -0.00849 | 0.00042711 | -19.88 | <.0001 |

这是对 y 与 x 、 z 之间进行多重线性回归分析的结果。 $F = 3589.84$ ， $P < 0.001$ ，说明拟合的模型有统计学意义。相关指数 $R^2 = 0.9985$ ，说明模型的拟合效果比较好。截距项 (Intercept)、 x 项系数、 z 项系数与 0 的差异均有统计学意义，对应的 t 值分别为 -11.06 、 26.14 和

-19.88, 而 P 值均小于 0.0001, 说明截距项和变量 x 、 z 均应保留在模型中。故拟合的方程为 $y = -22.07048 + 1.53801x - 0.00849z$ 则, 由于 $z = x^2$, 所以此二次项曲线模型应为 $y = -22.07048 + 1.53801x - 0.00849x^2$ 。

18.2 logistic 曲线回归分析

18.2.1 问题与数据

【例 18-2】某研究者欲分析某县疟疾发病的季节性特点, 观测了某县 1961—1996 年疟疾的月累计发病率(1/10 万), 结果见表 18-2。试对该资料进行曲线拟合。

表 18-2 某县疟疾月累计发病率

| 月份(月) | 累计发病率(1/10 万) |
|-------|---------------|
| 1 | 0.555 |
| 2 | 1.295 |
| ... | ... |
| 12 | 125.619 |

18.2.2 分析目的与统计分析方法的选择

通过后面的散点图(见图 18-2)可发现, 资料的趋势近似 S 形曲线, 故可采用 logistic 曲线拟合此资料。曲线上限为 130, 下限为 0, 上下渐近线之间的垂直间隔为 130。

18.2.3 SAS 程序

SAS 程序名为 sastjfx18_2.SAS。

```

%let ul=130; /* 1 */
data sastjfx18_2; /* 2 */
  do x=1 to 12;
    input y @@;
    z=log((&ul-y)/y);
    output;
  end;
  cards;

    0.555    1.295    2.751    11.116
    24.879   43.476   69.297   97.037
    114.631  121.645  124.412  125.619
  ;
run;
ods html;
proc gplot; /* 3 */
  plot y*x/haxis=0 to 12 vaxis=0
    to &ul by 10;
  symbol value=dot;
run;
proc reg data=sastjfx18_2
  outest=est(keep=Intercept x);
  model z=x; /* 4 */
run;quit;
data set1; /* 5 */
  set est;
  call
  symput('a1',exp(Intercept));
  call symput('b1',x);
run;

proc nlin data=sastjfx18_2; /* 6 */
  parms K=&a1 a=&a1 b=&b1;
  model y=K/(1+a*exp(b*x));
  output out=set2 p=yp r=resid;
run;
proc sql; /* 7 */
  create table set3(sum num,
  css num);
  insert into set3 select
  sum(resid**2),css(y) from set2;
quit;
data set4; /* 8 */
  set set3;
  r2=1-sum/css;
run;
proc print data=set2 round; /* 9 */
run;
proc print data=set4; /* 10 */
run;
ods html close;

```

程序中第 1 步设立宏变量 ul，表示曲线的上、下渐近线的垂直间距，其值为 130，以便于后面调用。第 2 步创建数据集，x 表示月份，y 表示疟疾月累计发病率，z 是对 y 进行 logit 变换所得的变量，以便后面进行曲线直线化分析。第 3 步是绘制散点图，以发现散点分布趋势，本资料散点趋势近似 S 形曲线。第 4 步对变量 z 和 x 进行直线回归分析，将所得直线的截距和斜率保存在数据集 est 中。第 5 步对数据集 est 进行操作，将以自然对数为底、以截距为指数的数值赋给宏变量 a1，将斜率赋给宏变量 b1。当然，也可直接将截距赋给宏变量 a1，将斜率赋给宏变量 b1，但第 6 步中 model 语句给出的曲线方程需改为“ $y = K / (1 + \exp(a + b * x))$ ”。第 6 步是调用 NLIN 过程来拟合 logistic 曲线，宏变量 ul、a1、b1 的值分别赋给变量 K、a、b 作为初值，并将所得曲线对各观测拟合的预测值及残差输出到数据集 set2 中，预测值以变量 yp 表示，残差以 resid 表示。第 7 步将数据集 set2 的内容输出到 output 窗口。第 7 步是调用 sql 过程，建立一个表 set3，包括 sum 和 css 两个变量，然后将数据集 set2 中所有观测 resid 的平方和赋给 set3 中的 sum 变量，将 set2 中所有观测 y 变量的离均差平方和赋给 css。第 8 步对数据集 set3 进行操作，计算新的变量 r2，即相关指数。第 9 步将数据集 set2 的内容输出到 output 窗口，round 选项用来指定数据最多保留 2 位小数。第 10 步将数据集 set4 的内容输出到 output 窗口。

18.2.4 主要分析结果及解释

图 18-2 是 SAS 程序绘制的散点图。可以看出，散点近似 S 形曲线趋势，上限为 130，下限为 0。

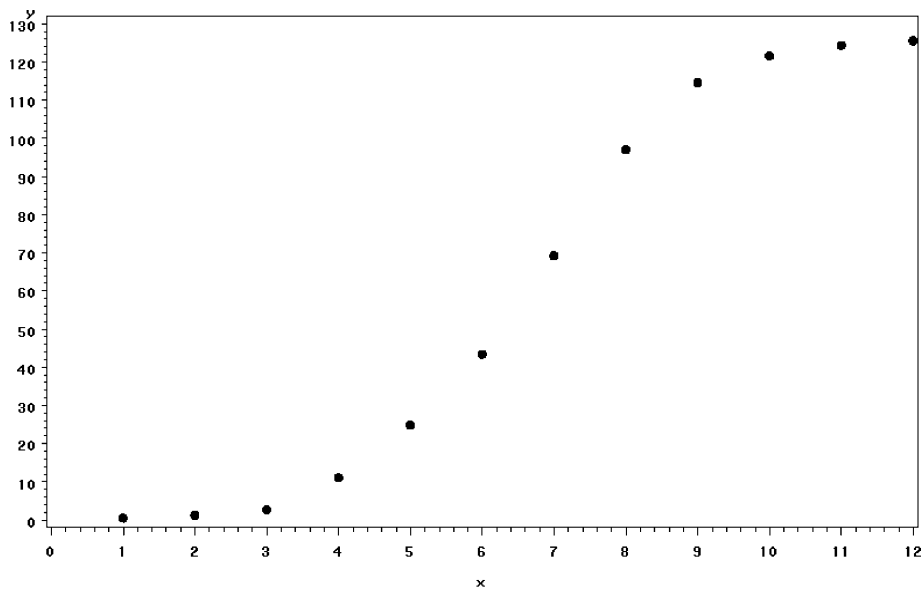


图 18-2 y 随 x 变化的散点图

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 6.00377 | 0.24283 | 24.72 | <.0001 |
| x | 1 | -0.84640 | 0.03299 | -25.65 | <.0001 |

这是曲线直线化分析的结果，即变量 z 与 x 之间直线回归分析的结果。截距为 6.00377，斜率为 -0.8464 ，它们与 0 之间的差异均有统计学意义，检验统计量 t 的值分别为 24.72、 -25.65 ，对应的 P 值均小于 0.0001。直线方程为 $z = 6.00377 - 0.8464x$ ，转换成 y 与 x 的关系，即 $y = 130 / (1 + e^{6.00377 - 0.8464x})$ 。

| Source | DF | Sum of Squares | Mean Square | F Value | Approx Pr > F |
|-------------------|----|----------------|-------------|---------|------------------|
| Model | 3 | 76037.4 | 25345.8 | 11770.1 | <.0001 |
| Error | 9 | 19.3807 | 2.1534 | | |
| Uncorrected Total | 12 | 76056.8 | | | |

| Parameter | Estimate | Approx Std Error | Approximate 95% Confidence Limits | |
|-----------|----------|---------------------|--------------------------------------|---------|
| K | 127.8 | 1.0737 | 125.4 | 130.2 |
| a | 394.6 | 65.1746 | 247.1 | 542.0 |
| b | -0.8879 | 0.0263 | -0.9474 | -0.8284 |

这是 NLIN 过程拟合模型的结果。对模型的检验结果为 $F = 11770.1$ ， $P < 0.0001$ ，说明所拟合的模型是有统计学意义的。模型的三个参数取值分别为： $K = 127.8$ 、 $a = 394.6$ 、 $b = -0.8879$ 。所以，本资料所拟合的 logistic 曲线方程为： $y = 127.8 / (1 + 394.6 \times e^{-0.8879x})$ 。

| | | | | | |
|-----|---------|----------|---------|--------|-------|
| Obs | x | y | z | yp | resid |
| 1 | 1 | 0.56 | 5.45 | 0.78 | -0.23 |
| 2 | 2 | 1.30 | 4.60 | 1.88 | -0.59 |
| ... | ... | ... | ... | ... | ... |
| 12 | 12 | 125.62 | -3.36 | 126.64 | -1.02 |
| Obs | sum | css | r2 | | |
| 1 | 19.3807 | 30827.91 | 0.99937 | | |

这是曲线拟合效果的结果。首先给出了曲线 $y = 127.8 / (1 + 394.6 \times e^{-0.8879x})$ 对各观测的拟合结果(为节省篇幅，中间几个观测予以省略)， yp 列为曲线对各观测的预测值， $resid$ 为预测值与观测值之间的残差。然后给出了对拟合整体效果的评价情况，残差平方和为 19.3807，相关指数 $R^2 = 0.99937$ ，说明此曲线拟合资料效果较好。

18.3 Gompertz 曲线回归分析

18.3.1 问题与数据

【例 18-3】 表 18-3 为人口简略寿命表，列出了 40 岁后不同年龄时的尚存人数。试拟合 Gompertz 回归曲线。

表 18-3 人口简略寿命表

| 年龄(岁) | 尚存人数(人) |
|-------|---------|
| 40 | 81277 |
| 45 | 79258 |
| ... | ... |
| 80 | 28074 |

18.3.2 分析目的与统计分析方法的选择

根据要求，需对本资料拟合 Gompertz 回归曲线。绘制散点图(见图 18-3)，找出曲线的上渐近线，初步判断，上渐近线约为 $y = 85000$ ，即 $L = 85000$ 。

Gompertz 曲线方程为:

$$y = Le^{-ae^{-bx}} \quad (18-1)$$

对式(18-1)进行两次对数变换,可得:

$$\ln\left(\ln\left(\frac{L}{y}\right)\right) = \ln(ae^{-bx}) = \ln(a) + \ln(e^{-bx}) = \ln a - bx \quad (18-2)$$

18.3.3 SAS 程序

SAS 程序名为 sastjfx18_3.SAS。

```
%let ul=85000; /* 1 */
data sastjfx18_3; /* 2 */
  do x=40 to 80 by 5;
    input y@@ ;
    z1 = &ul/y;
    z2 = log(z1);
    z = log(z2);
    output; end;
  cards;
81277 79258 76532
72850 67568 59911
50800 39325 28074
;
run;
ods html;
proc gplot; /* 3 */
  plot y*x;
run;
proc reg data=sastjfx18_3
  outest=est (keep=Intercept x);
  model z=x; /* 4 */
run;quit;
data set1; /* 5 */
  set est;
  call
  symput('a1',exp(Intercept));
  call symput('b1',-x);
run;
```

```
proc nlin data=sastjfx18_3; /* 6 */
  parms L=&ul a=&a1 b=&b1;
  model y=L* exp(-a* exp(-b* x));
  output out=set2 p=predicted
  r=resid;
run;
proc sql; /* 7 */
  create table set3 (sum num,
  css num);
  insert into set3 select
  sum(resid** 2),css(y) from set2;
quit;
data set4; /* 8 */
  set set3;
  r2=1-sum/css;
run;
proc print data=set2 round; /* 9 */
run;
proc print data=set4; /* 10 */
run;
ods html close;
```

第1步设立宏变量ul,作为上渐近线纵坐标值,根据观察,赋值为85000。第2步建立数据集,x为年龄,y为尚存人数,z是欲对Gompertz曲线进行直线化而求的新变量。第3步绘制散点图,观察y与x之间的变化趋势。第4步对z和x进行直线回归分析,并将所得的截距和斜率输出到数据集est中。第5步是对数据集est进行操作,将以e为底、以截距为指数的值赋给宏变量a1,将斜率的负值赋给宏变量b1,以便后面调用。第6步调用NLIN过程进行Gompertz曲线拟合,以宏变量ul、a1、b1的值分别赋给L、a、b作为初值,进行曲线拟合。将所得曲线对每个观测拟合的预测值和预测的残差输出到数据集set2中,分别记为predicted和resid。第7~10步的作用同程序sastjfx18_2,此处不再赘述。

18.3.4 主要分析结果及解释

图18-3是绘制的尚存人数与年龄之间变化趋势的散点图。在散点开始阶段,曲线变化比

较平稳，所以可确定其上渐近线的 y 值略大于 40 岁年龄组人数，可暂定为 85000。随着年龄的增加，尚存人数逐渐减少，当年龄增大到一定程度，其尚存人数将减至 0。

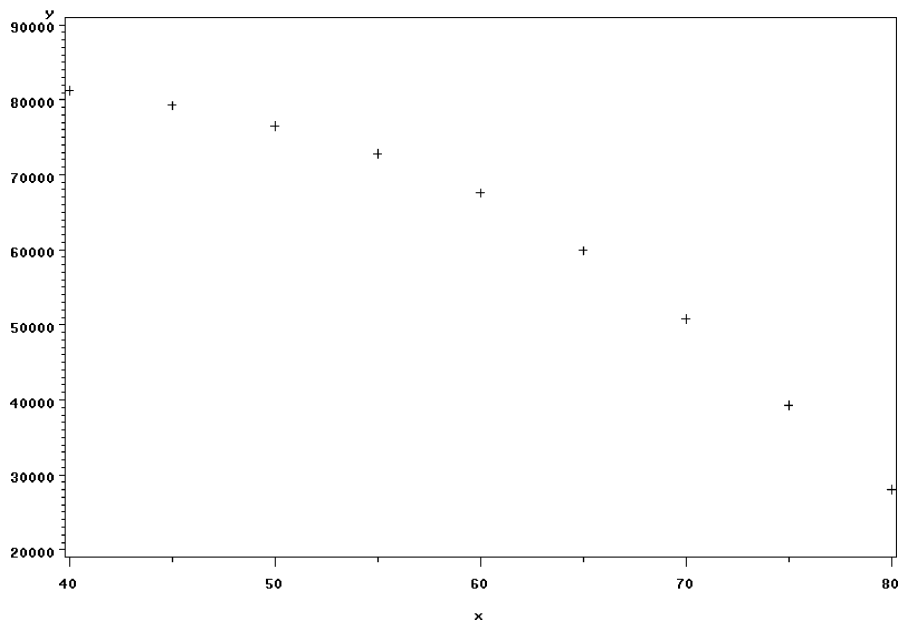


图 18-3 y 随 x 变化的散点图

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | -6.27667 | 0.03218 | -195.05 | < .0001 |
| x | 1 | 0.08010 | 0.00052433 | 152.77 | < .0001 |

这是对 z 与 x 进行直线回归分析的结果，截距和斜率分别为 -6.27667 和 0.0801 。

| Source | DF | Sum of Squares | Mean Square | F Value | Approx Pr > F |
|-------------------|----|----------------|-------------|---------|---------------|
| Model | 3 | 3.712E10 | 1.237E10 | 156257 | < .0001 |
| Error | 6 | 475135 | 79189.2 | | |
| Uncorrected Total | 9 | 3.712E10 | | | |

| Parameter | Estimate | Approx Std Error | Approximate 95% Confidence Limits | |
|-----------|----------|------------------|-----------------------------------|---------|
| L | 85679.2 | 383.8 | 84740.2 | 86618.3 |
| a | 0.00240 | 0.000229 | 0.00184 | 0.00296 |
| b | -0.0769 | 0.00121 | -0.0799 | -0.0740 |

| Approximate Correlation Matrix | | | |
|--------------------------------|-----------|-----------|-----------|
| | L | a | b |
| L | 1.0000000 | 0.9082606 | 0.8879742 |
| a | 0.9082606 | 1.0000000 | 0.9976098 |
| b | 0.8879742 | 0.9976098 | 1.0000000 |

这是 NLIN 过程拟合的结果。F = 156257，P < 0.0001，说明拟合的模型有统计学意义。三

个参数 L、a、b 的值分别为 85679.2、0.0024 和 -0.0769，所以 Gompertz 曲线模型为 $y = 85679.2 \times e^{0.0024e^{-0.0769x}}$ 。

| Obs | x | y | z1 | z2 | z | predicted | resid |
|-----|-----|-----------|--------------|------|---------|-----------|---------|
| 1 | 40 | 81277 | 1.05 | 0.04 | -3.11 | 81338.62 | -61.62 |
| 2 | 45 | 79258 | 1.07 | 0.07 | -2.66 | 79379.15 | -121.15 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 9 | 80 | 28074 | 3.03 | 1.11 | 0.10 | 27742.08 | 331.92 |
| | | | | | | | |
| | Obs | sum | css | | r2 | | |
| | 1 | 475135.43 | 2823635793.6 | | 0.99983 | | |

这是曲线拟合各观测的有关情况，predicted 为拟合的 Gompertz 曲线对各观测计算的预测值，resid 为测预测值与实际值之间的残差。sum 表示残差平方和，css 为变量 y 的离均差平方和。r2 表示拟合相关指数，其值为 0.99983，说明此 Gompertz 曲线拟合本资料效果很好。

18.4 二项型指数曲线回归分析

18.4.1 问题与数据

【例 18-4】 某双室模型药物静脉注射 100mg，测得各时间点的血药浓度结果如下：

t(时间, h): 0.165 0.500 1.000 1.500 3.000 5.000 7.500 10.000
C(血药浓度, μg/mL): 65.03 28.69 10.04 4.93 2.29 1.36 0.71 0.38
试拟合该资料的血药浓度 - 时间曲线图。

18.4.2 分析目的与统计分析方法的选择

资料中说明，药物是双室模型药物，且采用静脉注射，所以其“药 - 时”曲线应为二项型指数曲线。二项型指数曲线的曲线方程为：

$$y = A * e^{-\alpha * x} + B * e^{-\beta * x}$$
 (18-3)

18.4.3 SAS 程序

SAS 程序名为 sastjfx18_4.SAS。

```
data sastjfx18_4; /* 1 */
  input x y @@ ;
  z = log10(y);
  n = _n_;
  cards;
0.165 65.03 0.50 28.69
1.00 10.04 1.50 4.93
3.00 2.29 5.00 1.36
7.50 0.71 10.0 0.38
;
run;
proc gplot; /* 2 */
  plot z* x/haxis = 0 to 10
  vaxis = -0.5 to 2 by 0.5;

data d; /* 9 */
  set est2;
  call symput('a2',10** Intercept);
  call symput('b2', -2.303* x);
run;
proc nlin data = sastjfx18_4; /* 10 */
  parms l = &a2 a = &b2 m = &a1 b = &b1;
  bounds a > b > 0;
  model y = l* exp(- a* x) + m*
    exp(- b* x);
ods output EstSummary = data&i&j;
output out = cc&i&j
  p = predicted r = resid;
run;
```

```

symbol value = dot;
run;
%macro two(k);          /* 3 */
%do i = 3 %to &k - 3;
%do j = 3 %to &k - &i;
data a;                 /* 4 */
    set sastjfx18_4;
    where n >= &k - &i + 1;
run;
proc reg data = a
    outest = est (keep = Intercept x);
    model z = x;        /* 5 */
run;quit;
data b;                 /* 6 */
    set est;
    call symput('a1', 10 * Intercept);
    call symput('b1', -2.303 * x);
run;
data c;                 /* 7 */
    set sastjfx18_4;
    where n >= &k - &i - &j + 1
    and n <= &k - &i;
    y1 = &a1 * exp(-&b1 * x);
    yc = y - y1;
    zc = log10(yc);
run;
proc reg data = c
    outest = est2 (keep = Intercept x);
    model zc = x;       /* 8 */
run;quit;

data a&i&j;             /* 11 */
    set data&i&j (where =
    (Label1 = 'Objective'));
    keep nvalue1;
    rename nvalue1 = comb&i&j;
run;
proc transpose data = a&i&j
    out = b&i&j;         /* 12 */
run;
data kk;                /* 13 */
    set kk b&i&j;
run;
%end;
%end;
%mend two;              /* 14 */
data kk;                /* 15 */
    input _NAME_ $ coll;
    cards;
run;
%two(8)                 /* 16 */
ods html;               /* 17 */
proc print;
run;
ods html close;

```

程序第 1 步建立数据集, x 代表药物注射后时间, y 代表血药浓度, z 是对 y 进行以 10 为底的对数变换后的变量, n 是观测的序号。第 2 步绘制散点图, x 轴坐标为 0 ~ 10, 间隔为 1, z 轴为 -0.5 ~ 2, 间隔为 0.5。第 3 步建立宏程序, 宏名称为 `two`, 有一个宏参数 k , 其值为散点个数。在计算指数项参数的值时, 所得回归直线的斜率和截距对参数值的最终确定有重要影响。而回归直线的斜率和截距依赖于散点的选择, 所以在不同计算阶段, 选择合适的散点个数尤为重要。为寻找最优的散点组合, 可将不同计算阶段各种可能选取的散点个数都考虑进去, 由计算所得的截距和斜率推导出指数项参数的值, 进而代入 NLIN 过程作为初值, 拟合合适的曲线模型, 选取模型的标准是残差平方和最小。此宏以两个 `do` 循环语句开头, 目的就是将所有可能的选点组合挑选出来, 来找出残差平方和最小的曲线模型。第 4 步选取散点, 散点个数为 i 个, 保存在数据集 a 中。第 5 步基于数据集 a 中的观测, 拟合两个变量 x 和 z 的直线回归方程, 求得截距和斜率, 输出到数据集 est 中。第 6 步根据 est 中两个变量的值计算式 (18-3) 中指数项 $B * e^{-\beta * x}$ 的两个参数的值, 并赋给宏变量 $a1$ 和 $b1$ 。第 7 步仍然选取散点, 个数为 j 个, 并计算这 j 个点所在时间对应的残数浓度 yc 及其对数残数浓度 zc , 保存在数据集 c 中。第 8 步以数据集 c 中的观测为依据, 拟合对数残数浓度 zc 和药物注射后时间 x 之间的直线回归方程, 将其所得的截距和斜率保存在数据集 $est2$ 中。第 9 步根据 $est2$ 中两个变量的值计算式 (18-3) 指数项 $A * e^{-\alpha * x}$ 的两个参数的值, 并赋给宏变量 $a2$ 和 $b2$ 。第 10 步调用 NLIN 过程进行非线性曲线拟合, 以宏变量 $a1$ 、 $b1$ 、 $a2$ 、 $b2$ 的值作为二项型指数曲线 $y = l * e^{-ax} + m * e^{-bx}$ 中 m 、 b 、 l 、 a 的初值, 将参数估计的有关信息保存在数据集中, 数据集的命名规则是 `data` 后接两

个数字，这两个数字分别是宏变量 i 和 j 在循环过程中的取值。并将曲线方程对数据集各观测的拟合情况(含预测值及残差)输出到数据集 cc&i&j 中。第 11 步对数据集 data&i&j 进行操作，仅取数据集中变量 Label1 值为“Objective”的观测，保存到新数据集 a&i&j 中，新数据集中仅存有前述那个观测在 nvalue1 变量上的取值，此值为模型的残差平方和，且将 nvalue1 变量名更改为 comb&i&j，以利于后面合并数据集，将多次循环(即不同散点个数的组合)所得模型的残差平方和列在一起。第 12 步对数据集 a&i&j 进行转置，结果保存在新数据集 b&i&j 中，此步是将原本横向不便展示的数据集转置为纵向。第 13 步将 b&i&j 数据集中的所有观测与数据集 kk 纵向合并，仍存入 kk 中。此举是将每次循环所拟合的曲线模型的残差平方和都保存在数据集 kk 中。数据集 kk 是在第 15 步定义的，是一个空数据集。第 14 步结束宏内的 do 循环，结束宏程序 two。第 16 步执行宏程序 two，并将参数 k 的值赋为 8。第 17 步将数据集 kk 的内容输出，从而找出最小的残差平方和，进而给出曲线模型的方程。

18.4.4 主要分析结果及解释

图 18-4 是 SAS 程序绘制的 lgy - x 散点图。可见，曲线尾端的 4 个散点基本呈直线趋势，第 1 步计算时，就可以这 4 个点来拟合直线方程。当然，本资料已采用宏程序将各种可能的散点个数组合都进行拟合，可不必考虑如何选取散点个数的组合。但在散点个数较多时，则不宜将所有散点个数可能的组合都进行拟合，因为组合情形太多，计算量太大。此时，就可依据半对数散点图来选取合适的散点(指基本呈直线趋势的若干个散点)个数去拟合直线回归方程，进而计算相应指数项的参数值。

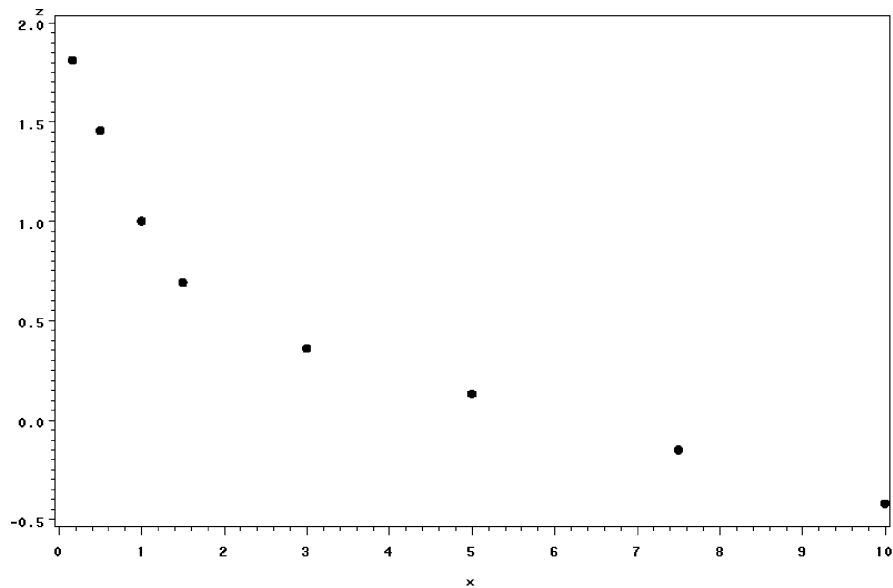


图 18-4 z 随 x 变化的散点图

| Obs | _NAME_ | col1 |
|-----|--------|----------|
| 1 | comb33 | 0.000945 |
| 2 | comb34 | 0.000945 |
| 3 | comb35 | 0.000945 |

| | | |
|---|--------|----------|
| 4 | comb43 | 0.000945 |
| 5 | comb44 | 0.000945 |
| 6 | comb53 | 0.000945 |

这是各种散点个数组合情况下, 所得到的曲线方程拟合资料时的残差平方和(见 col1 列)。因为是二项型指数曲线, 整个计算过程可分为两步, 每一步均需选取一次散点。本资料共有 8 个散点, 两计算阶段可能的散点个数组合有 6 种, 即上面所列的 6 种情形: 33、34、35、43、44、53。由 col1 列可以看出, 这 6 种散点个数组合, 最终所得到的曲线方程拟合本资料的残差平方和均为 0.000945。这里, 可以任选一种组合情形, 根据 NLIN 过程拟合的参数的值, 就可写出曲线方程了。以第 5 种散点个数组合情形为例(即 comb44), 在两个计算阶段均选取 4 个散点, 其分析结果如下:

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 0.69227 | 0.00542 | 127.75 | <.0001 |
| x | 1 | -0.11156 | 0.00078576 | -141.97 | <.0001 |

这是选取 $\lg y - x$ 散点图中后 4 个点拟合直线回归方程所得的结果。截距和斜率的值分别为 0.69227、-0.11156, 而 $10^{0.69227}$ 和 $-2.303 \times (-0.11156)$ 的值就是式(18-3)中第二个指数项 $B * e^{-\beta * x}$ 中 B 和 β 的值, 分别为 4.9234 和 0.2569。

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 1.97764 | 0.00213 | 927.28 | <.0001 |
| x | 1 | -1.18480 | 0.00227 | -521.67 | <.0001 |

这是选取 $\lg y_{残} - x$ 散点图中的 4 个点拟合直线回归方程所得的结果。截距和斜率的值分别为 1.97764、-1.1848, 而 $10^{1.97764}$ 和 $-2.303 \times (-1.1848)$ 的值就是式(18-3)中第一个指数项 $A * e^{-\alpha * x}$ 中 A 和 α 的值, 分别为 94.9814 和 2.7286。

| Source | DF | Sum of Squares | Mean Square | F Value | Approx Pr > F |
|-------------------|----|----------------|-------------|---------|---------------|
| Model | 4 | 5184.9 | 1296.2 | 5486551 | <.0001 |
| Error | 4 | 0.000945 | 0.000236 | | |
| Uncorrected Total | 8 | 5184.9 | | | |

| Parameter | Estimate | Approx Std Error | Approximate 95% Confidence Limits | |
|-----------|----------|------------------|-----------------------------------|---------|
| l | 94.3687 | 0.0436 | 94.2476 | 94.4899 |
| a | 2.7010 | 0.00319 | 2.6922 | 2.7099 |
| m | 4.7954 | 0.0395 | 4.6856 | 4.9052 |
| b | 0.2525 | 0.00252 | 0.2455 | 0.2595 |

Approximate Correlation Matrix

| | l | a | m | b |
|---|-----------|-----------|------------|------------|
| l | 1.0000000 | 0.1994946 | -0.2916815 | -0.3516155 |
| a | 0.1994946 | 1.0000000 | 0.8357829 | 0.6385723 |

| | | | | |
|---|------------|-----------|-----------|-----------|
| m | -0.2916815 | 0.8357829 | 1.0000000 | 0.8717299 |
| b | -0.3516155 | 0.6385723 | 0.8717299 | 1.0000000 |

这是 NLIN 过程拟合资料的结果。F = 5486551, P < 0.0001, 说明此曲线模型拟合资料有统计学意义。参数 l、a、m、b 的值分别为 94.3687、2.7010、4.7954 和 0.2525, 所以, 此二项型指数曲线方程为: $y = 94.3687 \times e^{-2.7010x} + 4.7954 \times e^{-0.2525x}$ 。得出曲线方程后, 若想得出曲线对各观测的拟合情况及曲线方程拟合资料的相关指数值, 可借助以下程序完成。当然, 这里需要调用程序 sastjfx18_4.sas 产生的数据集 cc44(cc 开头的其他数据集也可, 因各种散点个数组合所得到的曲线拟合本资料残差平方和相等, 否则, cc 后应接残差平方和最小的散点个数组合)。

```
proc sql;
    create table set1(sum num,css num);
    insert into set1 select sum(resid** 2),css(y) from cc44;
quit;
data set2;
    set set1;
    r2=1-sum/css;
run;
ods html;
proc print data=cc44 round; run;
proc print data=set2; run;
ods html close;
```

结果如下:

| Obs | x | y | z | n | predicted | resid |
|-----|-------|-------|-------|-----|-----------|-------|
| 1 | 0.17 | 65.03 | 1.81 | 1 | 65.03 | 0.00 |
| 2 | 0.50 | 28.69 | 1.46 | 2 | 28.68 | 0.01 |
| ... | ... | ... | ... | ... | ... | ... |
| 8 | 10.00 | 0.38 | -0.42 | 8 | 0.38 | 0.00 |

| Obs | sum | css | r2 |
|-----|------------|---------|---------|
| 1 | .000945014 | 3576.57 | 1.00000 |

这是曲线方程 $y = 94.3687 \times e^{-2.7010x} + 4.7954 \times e^{-0.2525x}$ 拟合资料的有关情况, 先给出了本曲线拟合各观测的预测值(predicted)和残差(resid), 后给出了曲线拟合资料整体的相关指数, 可见 $R^2 = 1.0000$, 说明此二项型指数曲线拟合资料效果很好。

需要说明的是, 并非所有的散点组合都是可行的。因为选取散点准备拟合回归直线时, 还需计算某些变量的对数值。若选点不合适, 则这些变量取值可能为负, 这样其对数值就无法计算了, 后续的结果也就不准确了。此时, 不适合以宏的方式来选取所有散点组合进行相应计算, 可根据散点趋势进行人工选点。

18.5 三项型指数曲线回归分析

18.5.1 问题与数据

【例 18-5】 某双室模型药物口服后, 测得不同时间的血药浓度如下:

| | | | | | | | | |
|-----------------------------|-----|------|------|------|------|------|------|------|
| t(h): | 0.1 | 0.3 | 0.5 | 1.0 | 2.5 | 5.0 | 7.5 | 10.0 |
| C(血药浓度, $\mu\text{g/mL}$): | 4.7 | 13.2 | 20.8 | 36.3 | 61.4 | 68.1 | 61.1 | 52.1 |

t(h): 15.0 20.0 25.0 30.0 40.0 50.0 60.0

C(血药浓度, $\mu\text{g/mL}$): 37.3 27.5 21.1 16.9 11.4 8.2 5.9

试拟合药-时曲线模型。

18.5.2 分析目的与统计分析方法的选择

双室模型药物口服后属血管外给药, 其在体内的“药-时”曲线是三项型指数曲线, 故对资料拟合三项型指数曲线。三项型指数曲线的曲线方程为:

$$y = Ne^{-k_a x} + Le^{-\alpha x} + Me^{-\beta x} \quad (18-4)$$

18.5.3 SAS 程序

SAS 程序名为 sastjfx18_5.SAS。

```
data sastjfx18_5;          /* 1 */
  input x y @@ ;
  z = log10(y);
  n = _n_;
  cards;
0.1  4.7  0.3 13.2  0.5 20.8
1.0 36.3  2.5 61.4  5.0 68.1
7.5 61.1 10.0 52.1 15.0 37.3
20.0 27.5 25.0 21.1 30.0 16.9
40.0 11.4 50.0  8.2 60.0  5.9
;
run;
proc gplot data=sastjfx18_5; /* 2 */
  plot z* x;
run;
data b;                    /* 3 */
  set sastjfx18_5;
  where n >= 13;
run;
ods html;                  /* 4 */
proc reg data=b
  outest=est (keep=Intercept x);
  model z=x;
run;quit;
data c;                    /* 5 */
  set est;
  call symput('a1', 10** Intercept);
  call symput('b1', -2.303* x);
run;
data s1;                   /* 6 */
  set sastjfx18_5;
  where n <= 12;
  y1 = &a1* exp(-&b1* x);
  yc = y - y1;
  zc = log10(yc);
run;
proc gplot data=s1;        /* 7 */
  plot zc* x;
run;

data s2;                  /* 11 */
  set s1;
  where n <= 7;
  y2 = &a2* exp(-&b2* x);
  yc2 = y2 - yc;
  zc2 = log10(yc2);
run;
proc gplot data=s2;        /* 12 */
  plot zc2* x;
run;
data f;                   /* 13 */
  set s2;
  where n <= 7;
run;
proc reg data=f outest
  =est3 (keep=Intercept x);
  model zc2=x;              /* 14 */
run;quit;
data g;                   /* 15 */
  set est3;
  call
  symput('a3', -10** Intercept);
  call symput('b3', -2.303* x);
run;
proc nlin data=sastjfx18_5; /* 16 */
  parms l = &a3 a = &b3 m = &a2 b = &b2
  k = &a1 c = &b1;
  model
  y = l* exp(-a* x) + m* exp(-b* x) +
  k* exp(-c* x);
  bounds a > b > c > 0;
  output out=pred p=predicted
  r=resid;
run;
proc sql;                 /* 17 */
  create table set1 (sum num,
  css num);
  insert into set1 select
  sum(resid** 2), css(y) from pred;
```

```

data d;                                /* 8 */
  set s1;
  where n >= 8;
run;
proc reg data = d
  outest = est2 (keep = Intercept x);
  model zc = x;                          /* 9 */
run;
data e;                                /* 10 */
  set est2;
  call symput ('a2 ', 10 * * Intercept);
  call symput ('b2 ', -2.303 * x);
run;
quit;
data set2;                             /* 18 */
  set set1;
  r2 = 1 - sum/css;
run;
ods html;
proc print data = pred round; /* 19 */
run;
proc print data = set2;          /* 20 */
run;
ods html close;

```

第1步建立数据集， x 表示服药后时间， y 表示服药后血药浓度， z 是对 y 进行以 10 为底的对数变换， n 表示观测号。第2步绘制半对数散点图，即以数据集 `sastjfx18_5` 中变量 x 为横轴，以 z 为纵轴，绘制 $\lg y - x$ 半对数散点图。由散点图可以看出，尾端 3 个点近似成一条直线。第3步选取数据集 `sastjfx18_5` 中的 $n = 13、14、15$ 的观测，即最后 3 个观测，保存至数据集 `b` 中。第4步对数据集 `b` 中的两个变量 z 和 x 进行直线回归分析，并将所得回归直线的截距和斜率保存至数据集 `est` 中。第5步对 `est` 数据集集中的截距和斜率进行相关计算，将所得的结果分别赋给宏变量 `a1` 和 `b1`。第6步选取数据集 `sastjfx18_5` 中 $n = 1$ 到 12 的 12 个观测。根据之前所得宏变量 `a1` 和 `b1` 的值来计算式 (18-4) 中 $M \times e^{-\beta x}$ 的值，即这 12 个点的外推浓度，记为 y_1 。以实际浓度 y 减去外推浓度 y_1 ，可得第一残数浓度，记为 y_c ，对其进行以 10 为底的对数变换，可得 y_c 的半对数浓度，记为 z_c 。所有观测及变量保存至数据集 `s1` 中。第7步对数据集 `s1` 中的变量 z_c 和 x 绘制散点图，以 x 为横轴， z_c 为纵轴，所得散点图为 $\lg y_{残1} - x$ 散点图（图中仅有 8 个散点，因第 1~4 个观测的 y_c 值为负，无法计算 z_c 值）。查看散点图可知，从右向左数，后 5 个点（即第 8~12 个观测）近似成一条直线。第8步选取数据集 `s1` 中 n 大于等于 8 的观测，保存在数据集 `d` 中。第9步以数据集 `d` 中的 5 个观测为基础，对变量 z_c 和 x 进行直线回归分析，并将所得回归直线的截距和斜率保存至数据集 `est2` 中。第10步对 `est2` 数据集集中的截距和斜率进行相关计算，将所得的结果分别赋给宏变量 `a2` 和 `b2`。第11步选取数据集 `s1` 中 $n = 1$ 到 7 的 7 个观测，根据之前所得宏变量 `a2` 和 `b2` 的值来计算式 (18-4) 中 $L \times e^{-\alpha x}$ 的值，记为 y_2 ，其含义为第一条残数线的外推浓度，将 y_2 减去第一残数浓度 y_c ，可得第二残数浓度，记为 y_{c2} ，对其进行以 10 为底的对数变换，可得 y_{c2} 的半对数浓度，记为 z_{c2} 。所有观测及变量保存至数据集 `s2` 中。第12步对数据集 `s2` 中的变量 z_{c2} 和 x 绘制散点图，以 x 为横轴， z_{c2} 为纵轴，所得散点图为 $\lg y_{残2} - x$ 散点图。查看散点图可知，这 7 个点近似成一条直线。第13步选取数据集 `s2` 中 n 小于等于 7 的所有观测，保存在数据集 `f` 中。其实，`s2` 中只有 7 个观测，本步骤此处略嫌多余，只需将第14步中的首句“`proc reg data = f`”改为“`proc reg data = s2`”，本步即可省略。但当 `s2` 中仅尾端部分点近似成一条直线时，可通过此步选取尾端呈直线趋势的那些点，以便下一步拟合直线回归方程。第14步对数据集 `f` 中的两个变量 z_{c2} 和 x 进行直线回归分析，并将所得回归直线的截距和斜率保存至数据集 `est3` 中。第15步对 `est3` 数据集集中的截距和斜率进行相关计算，将所得的结果分别赋给宏变量 `a3` 和 `b3`。第16步是调用 `NLIN` 过程对本资料拟合三项型指数曲线，以 `a1`、`b1`、`a2`、`b2`、`a3`、`b3` 分别作为三项型指数曲线 3 个指数项的 6 个参数 k 、 c 、 m 、 b 、 l 、 a 的初值。model 语句说明 y

与 x 的关系式为 $y = l \times e^{-ax} + m \times e^{-bx} + k \times e^{-cx}$ 。bounds 语句规定三个参数 a 、 b 、 c 的取值范围， $a > b > c > 0$ ，这是根据专业知识得到的。一般而言，血管外途径给药剂型在体内的吸收速度 a 大于分布速度 b ，分布速度 b 大于消除速度 c ，且三者均大于 0。output 语句是将所得曲线对各观测血药浓度 y 的预测值 predicted 及实际值与预测值之间的残差 resid 输出到数据集 pred 中。第 17 步借助 sql 过程创建数据集 set1，并将由 pred 数据集计算出来的残差平方和与血药浓度 y 的离均差平方和分别输出至数据集 set1 中的 sum 和 css 两变量列。第 18 步基于数据集 set1 计算相关指数 R^2 的值，以 r2 表示，并将新结果存入数据集 set2 中。第 19 步将数据集 pred 的内容输出到 output 窗口，round 选项用来指定结果最多保留 2 位小数。第 20 步将数据集 set2 的内容输出到 output 窗口。

18.5.4 主要分析结果及解释

图 18-5 是对资料绘制的 $\lg y - x$ 半对数散点图。可以看出，尾端 3 个点基本呈直线趋势，故可以这 3 个点来拟合直线回归方程。

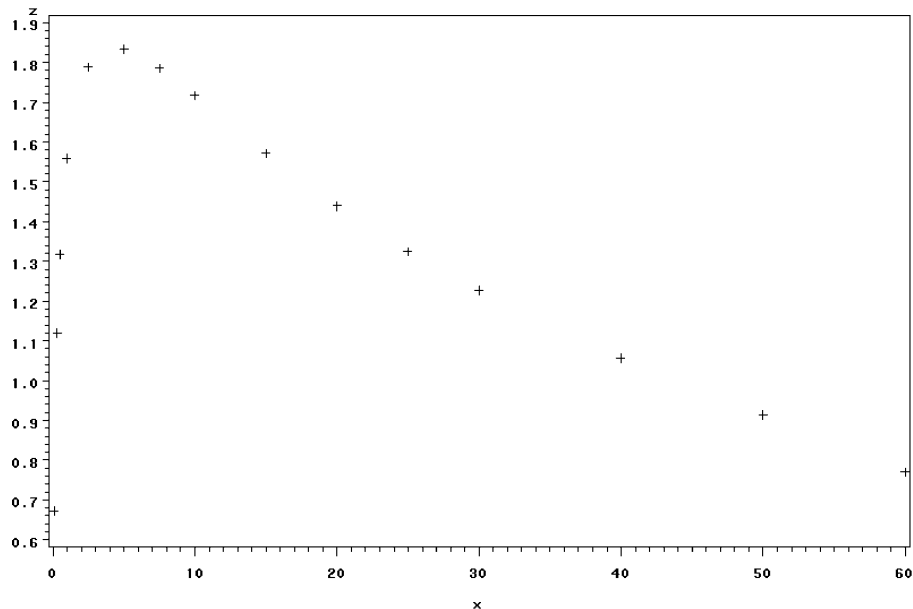


图 18-5 z 随 x 变化的散点图

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 1.62899 | 0.00018889 | 8623.87 | <.0001 |
| x | 1 | -0.01430 | 0.00000373 | -3836.1 | 0.0002 |

这是选取 $\lg y - x$ 散点图中尾端 3 个点拟合直线回归方程所得的结果。截距和斜率的值分别为 1.62899、-0.01430，而 $10^{1.62899}$ 和 $-2.303 \times (-0.01430)$ 的值就是式 (18-4) 中第三个指数项 $M \times e^{-\beta \times x}$ 中 M 和 β 的值，分别为 42.5588 和 0.0329。

图 18-6 是对资料绘制的 $\lg y_{残1} - x$ 半对数散点图。可以看出，尾端 5 个点基本成直线趋势，故可以这 5 个点来拟合直线回归方程。

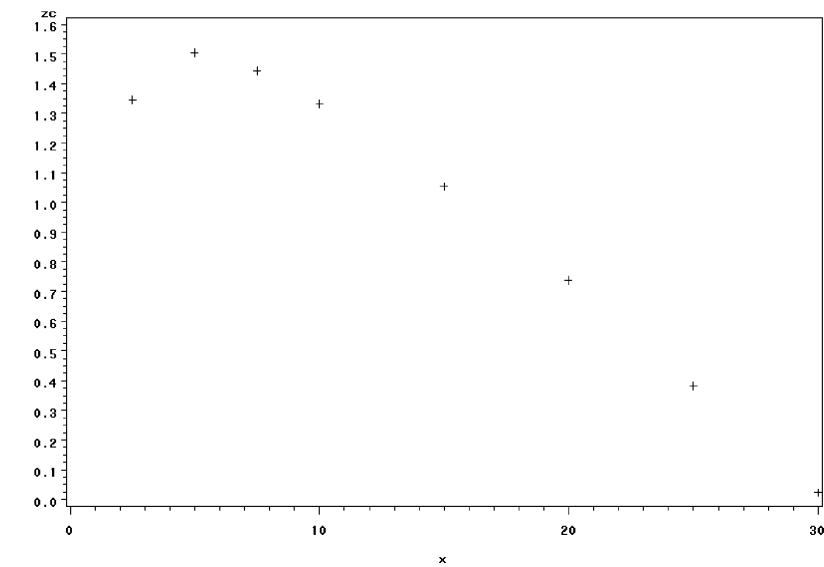


图 18-6 zc 随 x 变化的散点图

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 2.02117 | 0.04286 | 47.16 | <.0001 |
| x | 1 | -0.06573 | 0.00202 | -32.53 | <.0001 |

这是选取 $\lg y_{\text{残}1} - x$ 散点图中尾端 5 个点拟合直线回归方程所得的结果。截距和斜率的值分别为 2.02117、-0.06573，而 $10^{2.02117}$ 和 $-2.303 \times (-0.06573)$ 的值就是式 (18-4) 中第二个指数项 $L * e^{-\alpha * x}$ 中 L 和 α 的值，分别为 105.0 和 0.1514。

图 18-7 是对资料绘制的 $\lg y_{\text{残}2} - x$ 半对数散点图。可以看出，这 7 个点基本呈直线趋势，故可以这 7 个点来拟合直线回归方程。

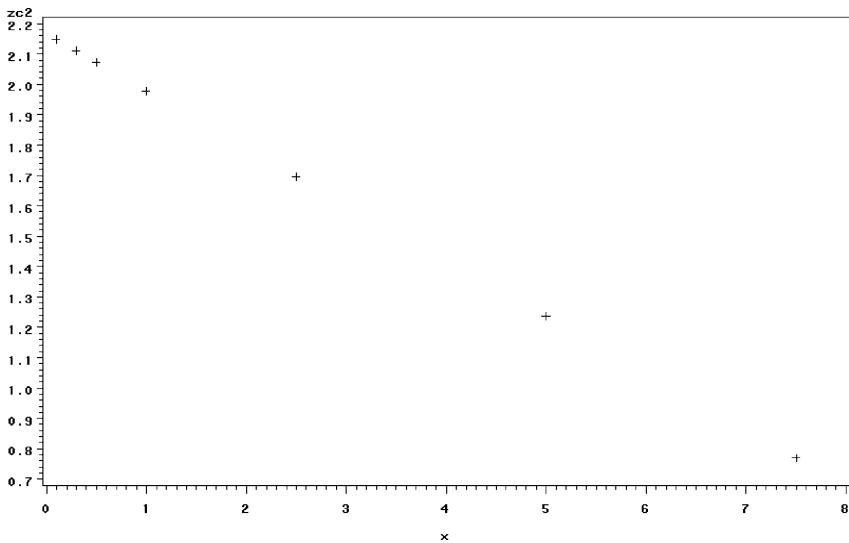


图 18-7 zc2 随 x 变化的散点图

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 2.16619 | 0.00125 | 1739.62 | <.0001 |
| x | 1 | -0.18625 | 0.00034951 | -532.88 | <.0001 |

这是选取 $\lg y_{\text{残}2} - x$ 散点图中全部 7 个点拟合直线回归方程所得的结果。截距和斜率的值分别为 2.16619、-0.18625，而 $-10^{2.16619}$ 和 $-2.303 \times (-0.18625)$ 的值就是式 (18-4) 中第三个指数项 $N * e^{-k_a * x}$ 中 N 和 k_a 的值，分别为 -146.6 和 0.4289。

| Source | DF | Sum of Squares | Mean Square | F Value | Approx Pr > F |
|-------------------|----|----------------|-------------|---------|---------------|
| Model | 6 | 19912.2 | 3318.7 | 1139046 | <.0001 |
| Error | 9 | 0.0262 | 0.00291 | | |
| Uncorrected Total | 15 | 19912.2 | | | |

| Parameter | Estimate | Approx Std Error | Approximate 95% Confidence Limits | |
|-----------|----------|------------------|-----------------------------------|---------|
| l | -124.6 | 0.9374 | -126.7 | -122.5 |
| a | 0.4793 | 0.00254 | 0.4735 | 0.4850 |
| m | 89.4276 | 0.5411 | 88.2036 | 90.6516 |
| b | 0.1197 | 0.00201 | 0.1152 | 0.1243 |
| k | 35.1908 | 0.9746 | 32.9861 | 37.3955 |
| c | 0.0298 | 0.000534 | 0.0286 | 0.0310 |

| Approximate Correlation Matrix | | | | | | |
|--------------------------------|------------|------------|------------|------------|------------|------------|
| | l | a | m | b | k | c |
| l | 1.0000000 | 0.9509332 | -0.2405154 | -0.9559101 | -0.8476178 | -0.7877144 |
| a | 0.9509332 | 1.0000000 | -0.3778639 | -0.8630504 | -0.7340173 | -0.6754606 |
| m | -0.2405154 | -0.3778639 | 1.0000000 | -0.0443234 | -0.3092117 | -0.3931607 |
| b | -0.9559101 | -0.8630504 | -0.0443234 | 1.0000000 | 0.9608652 | 0.9178828 |
| k | -0.8476178 | -0.7340173 | -0.3092117 | 0.9608652 | 1.0000000 | 0.9877205 |
| c | -0.7877144 | -0.6754606 | -0.3931607 | 0.9178828 | 0.9877205 | 1.0000000 |

这是 NLIN 过程拟合资料的结果。F = 1139046, P < 0.0001, 说明此曲线模型拟合资料有统计学意义。参数 l、a、m、b、k、c 的值分别为 -124.6、0.4793、89.4276、0.1197、35.1908 和 0.0298，所以，此三项型指数曲线方程为： $y = -124.6 \times e^{-0.4793x} + 89.4276 \times e^{-0.1197x} + 35.1908 \times e^{-0.0298x}$ 。

| Obs | x | y | z | n | predicted | resid |
|-----|------|------|------|-----|-----------|-------|
| 1 | 0.1 | 4.7 | 0.67 | 1 | 4.66 | 0.04 |
| 2 | 0.3 | 13.2 | 1.12 | 2 | 13.22 | -0.02 |
| ... | ... | ... | ... | ... | ... | ... |
| 15 | 60.0 | 5.9 | 0.77 | 15 | 5.94 | -0.04 |

| Obs | sum | css | r2 |
|-----|----------|---------|---------|
| 1 | 0.026222 | 6651.15 | 1.00000 |

这是所得曲线对各观测的拟合情况及曲线方程拟合资料的相关指数值， $R^2 = 1$ ，说明曲线 $y = -124.6 \times e^{-0.4793x} + 89.4276 \times e^{-0.1197x} + 35.1908 \times e^{-0.0298x}$ 对资料的拟合效果比较好。

18.6 本章小结

本章主要围绕生物医学科研中可能用到的较为复杂的回归曲线进行讲解,阐述了各种曲线的特点、适用情形及曲线拟合的实际步骤和 SAS 软件实现方法。其中,多项式曲线回归和 logistic 曲线回归简单一些,故计算和推理介绍较少;Gompertz 曲线回归和多项型指数曲线回归稍复杂一些,本章对它们的原理、公式推导也介绍得比较详细一些。

SAS 软件 NLIN 过程在拟合非线性趋势的曲线方面有良好的表现,但此过程对曲线模型参数初值的依赖较为严重,若初值赋值不甚合理,则可能无法得到收敛的模型或所得曲线模型仅是很小范围内的最优曲线。所以,初值的指定意义重大。对于复杂的曲线回归分析,若能进行直线化变换,可首先通过曲线直线化的方式求出曲线的方程,以此方程中各参数的值作为初值,借助 SAS 软件 NLIN 过程进行相应迭代,从而得到拟合效果更好的曲线方程;若无法进行直线化变换,则可根据曲线特点和相关公式,计算出参数的初值,然后以 NLIN 过程进行相应迭代,得出拟合效果较好的曲线方程。

另外,如何判断曲线方程对资料的拟合是否合适呢?这可从以下几个方面来把握:

(1) 所得曲线在专业上能得到适当的解释,变量系数的大小、正负不能与其专业含义相违背。

(2) 所得曲线应与观测点趋势一致。

(3) 对散点图趋势不甚明确的资料,应多拟合几种可能的曲线模型,从中选出最优的曲线。

总的来说,对于复杂回归曲线的分析,应着重把握专业依据和曲线特点两个方面,将两者结合起来,去探求符合专业知识、拟合资料效果良好的曲线模型。

(高辉 李子建)

第 19 章 多重线性回归分析

多重线性回归是指因变量为一个、自变量为多个的线性回归分析。如果根据专业知识得知，有多个变量分别取不同值，将会对某一个变量 Y 的取值有影响，且变量 Y 的取值是一个近似服从正态分布的连续型随机变量，自变量是一系列互相独立的定量变量或定性变量时，就可进行多重线性回归分析。本章主要介绍如何运用 SAS 软件中 REG 过程进行多重线性回归分析。

19.1 问题、数据及统计分析方法的选择

19.1.1 问题与数据

【例 19-1】 有研究认为血清中低密度脂蛋白增高是引起动脉硬化的一个重要原因，现测量了 30 名被怀疑患有动脉硬化的就诊患者的载脂蛋白 A I、载脂蛋白 B、载脂蛋白 E、载脂蛋白 C、低密度脂蛋白中的胆固醇含量，资料见表 19-1。试分析四种载脂蛋白对低密度脂蛋白中胆固醇含量的影响。

表 19-1 30 名患者载脂蛋白和低密度脂蛋白中胆固醇含量的测量结果

| 编 号 | 载脂蛋白 A I (mg/dl) | 载脂蛋白 B (mg/dl) | 载脂蛋白 E (mg/dl) | 载脂蛋白 C (mg/dl) | 低密度脂蛋白 (mg/dl) |
|-----|---------------------|-------------------|-------------------|-------------------|-------------------|
| 1 | 173 | 106 | 7.0 | 14.7 | 137 |
| 2 | 139 | 132 | 6.4 | 17.8 | 162 |
| ... | ... | ... | ... | ... | ... |
| 28 | 170 | 127 | 8.4 | 24.7 | 135 |
| 29 | 173 | 123 | 8.7 | 19.0 | 188 |
| 30 | 132 | 131 | 13.8 | 29.2 | 122 |

19.1.2 对数据结构的分析

本例中 30 名被怀疑患有动脉硬化的就诊患者应被视为随机抽自相应总体的一个样本，对他们未进行任何分组，低密度脂蛋白中的胆固醇含量、载脂蛋白 A I、载脂蛋白 B、载脂蛋白 E 和载脂蛋白 C 都可被视为定量变量，故可认为这是一个单组设计五元定量资料。

当然，在拟采用多重线性回归分析处理资料时，其数据结构不一定必须是多元定量资料，只要结果变量是定量的(最好服从正态分布)即可，自变量也可以是二值变量、多值有序变量或多值名义变量。

19.1.3 分析目的与统计分析方法的选择

对于表 19-1 这样的资料，有多种不同的统计分析目的，因此，也就有多种不同的统计分析方法。如果研究的目的是想要定量地分析多个自变量对一个定量的因变量的影响，找出因

变量是如何随自变量变化而变化的数量依存关系,然后对自变量的作用进行评价,或者是在此基础上进一步对因变量做出预测和估计,就可以采用多重线性回归分析。

19.2 多重线性回归分析概述

多重线性回归在实际中应用较为广泛,在用它对数据进行分析时,需要满足一定的条件,它要求自变量 X_1, X_2, \dots, X_m 和因变量 Y 之间的关系为线性关系;各个观察值之间相互独立;每一组自变量 X_1, X_2, \dots, X_m 对应的 Y 服从正态分布,且各个正态分布的方差相等。

多重线性回归分析的主要内容及任务就是要求出相应模型中参数的估计值,并对参数进行假设检验,对自变量进行共线性诊断和对观测点进行异常点诊断,结合统计学知识和专业知识对拟合的回归方程进行合理解释与应用。

如果某一项研究中,一个观测结果 y 与其他 p 个变量 x_1, x_2, \dots, x_p 的内在联系是线性的,它的第 α 个观测数据表示为 $(y_\alpha; x_{\alpha 1}, x_{\alpha 2}, \dots, x_{\alpha p}) \alpha = 1, 2, 3, \dots, n$ 。那么这一组数据可以写成如下形式:

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_p x_{2p} + \varepsilon_2 \\ \dots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_p x_{np} + \varepsilon_n \end{cases} \quad (19-1)$$

其中, $\beta_0, \beta_1, \dots, \beta_p$ 是 $p+1$ 个待估计参数, x_1, x_2, \dots, x_p 是 p 个可以精确测量或可控制的一般变量, $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 是 n 个相互独立且服从同一正态分布 $N(0, \sigma^2)$ 的随机变量。这就是多重线性回归的数学模型。

如果用矩阵来表示多重线性回归模型则更加方便,用矩阵表示如下:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \quad (19-1)$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (19-3)$$

那么上述模型可以写成矩阵的形式:

$$Y = X\beta + \varepsilon \quad (19-4)$$

其中, ε 为误差向量,它表示模型的预测值与观察值之间的差异,并且 ε 的维数为 n ,它的分量是相互独立的。为了能够使得回归方程为最优,则要求误差向量为最小。即

$$Q = \sum_{\alpha=1}^n (y_\alpha - \hat{y}_\alpha)^2 = \sum_{\alpha=1}^n (y_\alpha - b_0 - b_1 x_{\alpha 1} - b_2 x_{\alpha 2} - \dots - b_p x_{\alpha p})^2 \quad (19-5)$$

最小。为了达到此目的,对向量 β 的估计则采用最小二乘估计,其计算公式为:

$$\beta = (X'X)^{-1}X'Y \quad (19-6)$$

其中, X 为 $n \times p$ 阶矩阵, X' 为 X 的转置, $(X'X)$ 为对称的 $p \times p$ 方阵, 通常 $X'X$ 称为设计矩阵, $(X'X)^{-1}$ 为 $X'X$ 的逆矩阵, Y 为因变量的 $n \times 1$ 向量, β 为待估参数, 即回归系数的 $p \times 1$ 向量, 这里的 n 为观察值组数, m 为待估计的回归系数的个数。由于各自变量的单位不同, 因而不能直接根据原始数据计算的结果来评价各自变量对因变量的贡献大小, 最好先对原数据进行标准化后再计算偏回归系数, 称为标准化偏回归系数。其值越大, 说明该自变量对因变量的作用越大。

由数理统计方法可以证明用最小二乘法对 β 进行的估计为无偏估计, 从而也体现了最小二乘估计的优良性质。但是采用此法时应注意一些问题: 一是 Y 的方差要一致, 即各次观测的误差要基本一致, 也就是说要方差齐性; 二是各次观测结果要相互独立。如果因变量 Y 的分布符合正态分布, 则可以对回归方程做假设检验或构造回归系数的置信区间, 因此因变量 Y 的正态性也是一个很重要的要求。在 SAS 统计软件中, 若在 MODEL 语句中不加选择项 “STB”, 仅给出未标准化的偏回归系数的估计值; 反之, 同时给出未标准化和标准化的偏回归系数的估计值。

在建立了回归方程以后, 就可以对因变量与自变量之间的线性依存关系进行定量的分析, 进而还可以利用回归方程对因变量进行预测, 或者是实现对于自变量的统计控制。

19.3 产生哑变量和派生变量的方法

19.3.1 如何产生哑变量

当自变量不是定量变量时, 应对其进行预处理。对于二值自变量直接赋 0、1 值。对于多值名义变量, 不能直接将其代入回归方程进行计算, 需要对其产生哑变量, 然后将哑变量纳入模型进行分析。对于多值有序变量, 最好根据其各等级对结果变量的影响情况赋相应的值, 但难度较大。习惯上采用按连续性变量那样处理, 直接将其代入方程, 但这样做不够科学, 既简便又比较稳妥的做法是将其视为多值名义变量, 产生哑变量。

假定自变量中有多值名义变量, 怎样对其产生哑变量呢? 现假定自变量中有一个自变量为血型 (blood), 对其产生哑变量的方法如下:

首先, 用 SAS 中的 FREQ 过程求出数据集中 A、B、AB、O 四种血型所对应的频数, 其 SAS 过程步为:

```
PROC FREQ; TABLES blood; RUN;
```

假定 O 型血的人数最多, 于是, 就可以 O 血型的人为基准, 产生三个哑变量 X_1 、 X_2 和 X_3 分别代表 A、B、AB 血型(注: 哑变量的个数为多值名义变量的水平数减一), 即

| Blood | X_1 | X_2 | X_3 |
|-------|-------|-------|-------|
| A | 1 | 0 | 0 |
| B | 0 | 1 | 0 |
| AB | 0 | 0 | 1 |
| O | 0 | 0 | 0 |

上述做法如何在 SAS 程序的数据步中体现出来呢? 在 INPUT 语句和 CARDS 语句之间加上下面的 SAS 语句即可:

```
X1 = 0; X2 = 0; X3 = 0;  
IF blood = 'A' THEN X1 = 1;  
ELSE IF blood = 'B' THEN X2 = 1;  
ELSE IF blood = 'AB' THEN X3 = 1;
```

于是，在写 REG 过程步的 MODEL 语句等号后面，不写 blood，写 X1 - X3 即可。

19.3.2 如何产生派生变量

研究者在收集资料时，根据基本常识和专业知识选定了一些最有可能影响因变量的自变量，但它们与因变量之间的关系并非一定是一次方的形式，有可能是二次方的形式或三次方的形式；也有可能某些自变量之间以连乘的形式与因变量建立联系。所以，在进行多重线性回归分析之前，最好通过探索性分析，了解自变量与因变量可能存在哪些特殊的联系方式，有的放矢地引入那些特殊形式的“新自变量”，它们被称为“派生变量”。恰当地引入某些派生变量，将有利于构建拟合度更高的多重线性回归模型。

19.4 自变量各种筛选方法介绍

19.4.1 筛选自变量的必要性

研究者事先选定的全部自变量不一定都对因变量有重要的影响作用，有必要通过假设检验将那些作用微弱的自变量淘汰掉，以提高多重线性回归模型的质量并获得精炼的多重回归模型。

19.4.2 筛选自变量的方法

在 SAS 软件中，有 8 种筛选自变量的方法，它们分别称为前进法、后退法、逐步法、最大 R^2 增量法、最小 R^2 增量法、 R^2 选择法、修正 R^2 选择法和 CP 统计量选择法，后三种可被统称为“最优回归子集法”。

1. 前进法

模型中自变量从无到有依次选一变量进入模型。引入变量时，选择偏回归平方和最大的自变量，根据偏回归平方和计算 F 统计量及 P 值，当 P 小于事先确定的显著性水平，也就是纳入标准时，该变量入选，否则不能入选。每一个自变量按照这样逐一引入回归方程，直到没有自变量可入选为止。该方法可以自动去掉高度相关的自变量。它的局限在于纳入标准取值小时，可能没有一个自变量能入选；纳入标准较大时，开始选入的自变量后来在新条件下不再进行检验，因而不能剔除后来变得无统计学意义的自变量。

2. 后退法

回归模型中最初包含所有的自变量，然后逐步剔除无统计学意义的自变量。每次进行剔除时，选择模型中偏回归平方和最小的自变量，计算 F 统计量和 P 值，当 P 值小于事先确定的剔除标准，则将此变量保留在方程中，否则，将其从模型中剔除。从 P 值最大的自变量开始逐一剔除，直到模型中没有变量可以被剔除时为止。后退法的局限在于剔除标准较大时，任何一个自变量都不能被剔除；剔除标准较小时，开始被剔除的自变量后来在新条件下即使变得对因变量有较大的贡献了，也不能再次被选入回归模型并参与检验。

3. 逐步筛选法

逐步筛选法可以看成前进法和后退法的结合。模型中的自变量从无到有像前进法那样, 根据 F 统计量和 P 值按纳入标准决定该自变量是否入选; 当模型选入自变量后, 又像后退法那样, 根据 F 统计量和 P 值按剔除标准剔除无统计学意义的自变量。对每一个自变量逐一进行这个过程, 直到没有自变量可入选, 也没有自变量可被剔除, 则停止逐步筛选过程。逐步筛选法能比前进法和后退法更好地选出自变量构造模型, 但也有它的局限性: 其一, 在有 p 个自变量入选后, 选第 $p+1$ 个自变量时, 对它来说, 前 p 个自变量不一定是最佳组合; 其二, 选入或剔除自变量仅以 F 值作为标准, 完全没考虑其他标准。

4. 最大 R^2 增量法

首先找到具有最大决定系数 R^2 的单变量回归模型, 接着引入产生最大 R^2 增量的另一自变量。然后对于该两变量的回归模型, 用其他自变量逐次替换, 每次计算回归方程的 R^2 , 如果换后的模型能产生最大 R^2 增量, 即为两变量最优回归模型, 如此再找下去, 直到入选自变量数太多, 使设计矩阵不再满秩时为止。该方法也是一种逐步筛选法, 只是筛选变量所用的准则不同, 不是用 F 值, 而是用决定系数 R^2 判定自变量是否入选。因它不受纳入和剔除标准的限制, 总能从变量中找到相对最优者, 这就克服了前述三种方法的局限性, 即找不到任何自变量可进入模型的情况。本法的局限与逐步筛选法相似, 当有 p 个变量入选后, 选第 $p+1$ 个变量时, 对它来说, 前 p 个变量不一定是最佳组合。

5. 最小 R^2 增量法

首先找到具有最小决定系数 R^2 的单变量回归模型, 然后从其余自变量中选出一个自变量, 使它构成的模型比其他自变量所产生的 R^2 增量都小, 不断用新变量来替换老变量, 依次类推, 这样就会顺次列出全部单变量回归方程, 依次为 R^2 最小者, R^2 增量最小者, R^2 增量次小者, \dots , R^2 增量最大者, 在这些含一个自变量的回归方程中, 最后一个为单变量最佳模型; 两变量最小 R^2 增量的筛选类似本节第 4 种方法, 但引入的是产生最小 R^2 增量的另一自变量。对该两变量的回归方程, 再用其他自变量替换, 换成产生最小 R^2 增量者, 直至 R^2 不能再增加, 即两变量最优回归方程。依次类推, 继续找含三个或更多自变量的最优回归模型, 变量有进有出。它与本节第 4 种方法得到的“最优”模型常常相同, 但它在寻找最优方程过程中所考虑的中间模型要比本节第 4 种方法多。本法的局限与本节第 3、4 种方法相似, 当有 p 个自变量入选后, 选第 $p+1$ 个变量时, 每次只有一个自变量进或出, 各自变量间有复杂关系时, 就有可能找不到最佳组合。

6. R^2 选择法

从各自变量所有可能子集中选出某个子集, 使该子集所构成的模型的决定系数 R^2 最大。本法和后面的第 7、8 两种方法分别是按不同标准选出回归模型自变量的最优子集。 R^2 选择法的局限在于计算量比较大, 特别是在自变量个数较多时。另外 R^2 总是随着自变量个数的增加而增大, 即使某个新增加的自变量经检验可能没有统计学意义。

7. 校正 R^2 选择法

根据校正 R^2 的取最大的原则, 从模型的所有自变量子集中选出规定数目的子集。程序能运行的条件是设计矩阵要满秩。

8. Mallow's C_p 选择法

Mallow's C_p 也可以用来评价回归模型，其计算公式为：

$$C_p = \frac{(SS_E)_p}{(MS_E)_m} - (n - 2p) \quad (19-7)$$

式中， p 为方程中包含的自变量个数， $(SS_E)_p$ 为包含 p 个自变量的回归方程所对应的残差平方和， $(MS_E)_m$ 为包含所有 m 个自变量的回归方程对应的残差均方。

该方法就是选择 C_p 最接近于 p 的回归方程为最优方程。

R^2 选择法、校正 R^2 选择法和 Mallow's C_p 选择法都属于最优回归子集法。

19.4.3 如何在多个多重线性回归方程中选择最佳者

对于究竟哪一种筛选变量的方法最好，没有绝对的定论。一般来说，逐步回归法和最优回归子集法较好。对于一个给定的资料，最好尝试多种变量筛选的方法，结合以下几条评价标准，从中选择最佳者。

- (1) 拟合的多重回归方程在整体上有统计学意义。
- (2) 多重回归方程中各回归参数估计值的假设检验结果都有统计学意义。
- (3) 多重回归方程中各回归参数估计值的正负号与专业上的含义相吻合。
- (4) 根据多重回归方程计算出因变量的所有预测值在专业上都有意义。
- (5) 若有多个较好的多重回归方程时，残差平方和较小且多重回归方程中所含的自变量的个数又较少者为最佳。

19.5 自变量各种共线性诊断方法介绍

19.5.1 共线性的概念

当一个或几个自变量可以由另外的自变量线性表示时，称该自变量与另外几个自变量间存在多重共线性关系。在实际研究中，经常会遇到自变量间存在线性关系或者接近线性关系，当自变量间有共线性关系时，必然使回归系数的估计变得不稳定。自变量间即使不是完全的共线性，而是近似的共线性时（如相关系数接近于1），将使最小二乘法失效，使得回归方程中参数变得不稳定，因此对回归系数的估计、预测值的精度产生很大的影响。所以，需要对自变量之间是否存在较强的共线性关系进行研究，称为共线性诊断。

19.5.2 共线性的诊断

1. 用条件数和方差分量进行共线性诊断

先求出信息矩阵 $X'X$ 的各特征根，条件指数定义为最大特征根与每个特征根比值的平方根，其中最大条件指数 k 称为矩阵 $X'X$ 的条件数。条件数大，说明设计矩阵有较强的共线性，结果可能会不稳定，甚至使离开试验点的各估计值或预测值毫无意义。直观上，条件数度量了信息矩阵 $X'X$ 的特征根散布程度，可用来判断多重共线性是否存在以及多重共线性严重程度。在应用中，若 $0 \leq k < 10$ ，则认为没有多重共线性； $10 \leq k \leq 30$ ，则认为存在中等程度或较强的多重共线性； $k > 30$ ，则认为存在严重的多重共线性。

强的多重共线性同时还会表现在变量的方差分量上,对条件数同时有两个以上自变量的方差分量超过 50%,就意味这些自变量间有一定程度的相关性。

2. 用方差膨胀因子进行共线性诊断

首先对容许度进行说明,对一个入选自变量而言,该统计量等于 $1 - R^2$,这里 R^2 是把该自变量作为因变量对模型中所有其余自变量作回归时的决定系数, R^2 越大,则容许度越小,表明该自变量与其他自变量之间的关系密切。

方差膨胀因子 VIF 被定义为容许度的倒数,对于不好的试验设计,VIF 的取值可能趋于无限大。VIF 达到什么数值就可认为自变量间存在共线性,目前尚无标准的临界值。有人根据经验得出:VIF > 10 时,就有严重的多重共线性存在。

19.6 各种异常点诊断方法介绍

19.6.1 异常点的概念

若个别观测点与多数观测点偏离很远,它们可能会对回归的估计以及其他推断产生很大的影响,对这一问题进行监测和分析的方法,称为异常点诊断。

19.6.2 异常点的诊断

1. 用学生化残差等统计量进行异常点诊断

对因变量的预测值影响特别大,甚至容易导致相反结论的观测点,被称为强影响点或异常点。可以用于异常点诊断的统计量很多,包括残差、学生化残差、杠杆率和 Cook's D 统计量等等,其中比较便于判断的是学生化残差统计量。当该统计量的绝对值大于 2 时,所对应的观测点可能是异常点,此时,需认真核对原始数据。若属抄写或输入数据时人为造成的错误,应当予以纠正;若属非过失误差所致,可将异常点剔除前后各作出一个最好的回归方程,并对所得到的结果进行分析和讨论。如果有可能,最好在此点上补做试验,以便进一步确认可疑的“异常点”是否确属异常点。

2. 用绘制残差图的方法进行异常点诊断

此外,还可以通过绘制残差图来对模型假设的合理性进行考察,残差图的纵坐标可以是残差、标准化残差或学生化残差,横坐标可以是因变量的估计值 \hat{Y} 或者某一自变量的值等。在残差图中,如果各个散点随机均匀地散布在直线 $y = 0$ 的上下两侧,说明资料符合模型的假设。如果呈现出某种特别的趋势,就要考虑因变量与自变量之间的关系可能非线性、方差不齐或者残差不独立这几种情况之一。

19.7 自变量作用大小的评价

在建立了回归方程之后,不能直接根据各自变量回归系数绝对值的大小来评价该自变量对因变量的作用大小,因为自变量的单位不尽相同,回归系数的大小要受到单位的影响。如果要比较各自变量的作用大小,应消除自变量单位的影响,这就要求标准化的回归系数。标准化回归系数没有量纲,统计学上常用它的绝对值大小来衡量自变量对因变量影响的相对重要性,标准化回归系数的绝对值越大,说明该自变量对因变量的作用越大。

19.8 多个多重回归方程优劣的评价

经过回归诊断和变量筛选后，是否一定能找到一个最好的回归方程呢？那很难说！对于究竟哪一种筛选变量的方法最好，没有绝对的定论。一般来说，逐步回归法和最优回归子集法较好。对于一个给定的资料，可试用多种变量筛选的方法，结合以下几条评价标准，从中选择最佳者。

- (1) 拟合的多重回归方程在整体上有统计学意义。
- (2) 多重回归方程中各回归参数的估计值的假设检验结果都有统计学意义。
- (3) 多重回归方程中各回归参数的估计值的正负号与其后的变量在专业上的含义相吻合。
- (4) 根据多重回归方程计算出因变量的所有预测值在专业上都有意义。
- (5) 若有多个较好的多重回归方程时，残差平方和较小且多重回归方程中所含的自变量的个数又较少者为最佳。
- (6) 若以上各条都满足的多重回归方程仍有一个以上，此时，可选择自变量的数值采集起来方便且误差小的那个多重回归方程为最终的回归方程。

19.9 多重线性回归分析的 SAS 实现

19.9.1 SAS 程序及说明

设对例 19-1 中资料拟合多重线性回归模型的程序名为 SASTJFX19_1.SAS，程序如下：

```
data SASTJFX19_1;
input id x1 -x4 y@@ ;
cards;
1 173 106 7.0 14.7 137
2 139 132 6.4 17.8 162
3 198 112 6.9 16.7 134
4 118 138 7.1 15.7 188
5 139 94 8.6 13.6 138
6 175 160 12.1 20.3 215
...
;
run;
```

```
ODS HTML STYLE = MINIMAL;
proc reg data = SASTJFX19_1;
model y = x1 - x4 / selection = stepwise;
run;

proc reg data = SASTJFX19_1;
model y = x2 x4 / noint collin;
collino int r stb;
plot r.* p.;
run;
ODS HTML CLOSE;
```

程序说明：首先建立数据集 SASTJFX19_1，变量 x1 - x4 分别表示四种载脂蛋白，y 代表低密度脂蛋白中的胆固醇含量。后面的两个 REG 过程步实现多重线性回归。第一个 REG 过程步的主要目的是进行变量筛选，model 语句中等号的左侧是因变量 y，等号右侧为自变量 x1 - x4。斜杠“/”后的选项用来定义分析过程中的具体要求，这里选项 selection 规定自变量筛选的方法，其取值 stepwise 表示使用逐步选择法。第二个 REG 过程步使用筛选得到的变量 x2 和 x4 进行建模；选项 noint 表示拟合的模型中不包含截距项，这是由于第一个 REG 过程中对截距项的检验没有统计学意义；选项 collin 和 collino int 实现共线性诊断，其中 collin 输出未对截距项进行调整的结果，collino int 输出将截距项调整出去以后的结果；选项 r 要求进行残差分析，输出实际值、预测值、残差、预测值和残差的标准误、学生化残差和 Cook's D 统计量；选项 stb

表示输出标准化回归系数。plot 语句用来绘制散点图, plot r. * p. 指定绘制残差图, 其中纵轴表示残差(r.), 横轴表示预测值(p.)。

19.9.2 主要分析结果及解释

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

| Step | Variable Entered | Variable Removed | Number Vars In | Partial R-Square | Model R-Square | C(p) | F Value | Pr > F |
|------|------------------|------------------|----------------|------------------|----------------|---------|---------|--------|
| 1 | x2 | | 1 | 0.3158 | 0.3158 | 13.2447 | 12.92 | 0.0012 |
| 2 | x4 | | 2 | 0.2219 | 0.5377 | 2.5168 | 12.96 | 0.0013 |

此部分结果由第一个 REG 过程输出, 主要是对自变量进行筛选, 筛选过程分为两个步骤, 由于结果较多, 这里省略了具体每一步的输出结果, 只给出最后总结部分的内容。结果中开头部分的文字说明了纳入和排除标准为 0.15, 这是系统默认的标准, 用户也可以通过 model 语句之后的 sle 和 sls 选项来定义其他标准。结果中的第一列代表筛选的各个步骤, 第二、三列给出了每一步中入选和剔除变量的名称, 第四列表示还留在方程中的变量个数, 第五列和第六列表示偏 R^2 与模型 R^2 , 第七列为 C_p 统计量, 第八列和第九列为对该变量进行检验的 F 值和 P 值。由这部分结果可以看出, 模型中最终保留了 x2 和 x4。在考察了其他一些筛选方法(结果从略), 进而综合考虑的基础上, 最终选择了 x2 和 x4 建立回归方程。

The REG Procedure

Dependent Variable: y

Number of Observations Read 30
Number of Observations Used 30

这是第二个 REG 过程输出的第一部分结果, 说明读入的观测数和使用的观测数, 这里都是 30 个。

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|-------------------|----|----------------|-------------|---------|--------|
| Model | 2 | 706744 | 353372 | 611.30 | <.0001 |
| Error | 28 | 16186 | 578.06940 | | |
| Uncorrected Total | 30 | 722930 | | | |

第二部分是对整个模型进行方差分析的结果, 包括模型和误差项对应的自由度、离均差平方和与均方, 以及 F 值和 P 值, 这里 $F = 611.30$, $P < 0.0001$, 说明所求的回归模型总体上来说有统计学意义。

| | | | |
|----------------|-----------|----------|--------|
| Root MSE | 24.04307 | R-Square | 0.9776 |
| Dependent Mean | 151.66667 | Adj R-Sq | 0.9760 |
| Coeff Var | 15.85258 | | |

第三部分是一些描述性统计量, Root MSE 为误差均方的平方根, 即剩余标准差; R-square 为决定系数; Dependent Mean 为因变量的均数; Adj R-Sq 为校正决定系数; Coeff Var 为因变量的变异系数。

| Parameter Estimates | | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|-----------------------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t | Standardized Estimate |
| x2 | 1 | 1.55505 | 0.10969 | 14.18 | <.0001 | 1.27013 |
| x4 | 1 | -2.17820 | 0.64754 | -3.36 | 0.0022 | -0.30138 |

第四部分是参数估计的结果，由左至右的各列依次为变量名、自由度、参数估计值、标准误、对各参数进行检验的 t 值和 P 值，以及标准回归系数，这一项是由选项 stb 产生的。由于规定了选项 noint，所以并没有输出截距项。变量 x2 和 x4 的回归系数分别为 1.55505 和 -2.17820，对其进行检验的 P 值都小于 0.01，说明这两者与 0 之间的差异都有统计学意义，从而可以写出回归方程为：

$$\hat{y} = 1.55505x_2 - 2.17820x_4$$

此外，由标准化回归系数可以看出，与变量 x_4 相比， x_2 对低密度脂蛋白中胆固醇含量的影响要大一些。

| Collinearity Diagnostics | | | | |
|--------------------------|------------|-----------------|-------------------------|---------|
| Number | Eigenvalue | Condition Index | Proportion of Variation | |
| | | | x2 | x4 |
| 1 | 1.94889 | 1.00000 | 0.02556 | 0.02556 |
| 2 | 0.05111 | 6.17477 | 0.97444 | 0.97444 |

| Collinearity Diagnostics (intercept adjusted) | | | | |
|---|------------|-----------------|-------------------------|---------|
| Number | Eigenvalue | Condition Index | Proportion of Variation | |
| | | | x2 | x4 |
| 1 | 1.94889 | 1.00000 | 0.02556 | 0.02556 |
| 2 | 0.05111 | 6.17477 | 0.97444 | 0.97444 |

第五部分是共线性诊断的结果，包括未对截距项进行调整以及将截距项调整出去之后的结果。本例中由于拟合的模型并不包含截距项，所以两者的结果是一致的。这部分结果具体包括特征根、条件指数与方差分量，由以上可以看出条件数为 6.17477，该值小于 10，说明这两个自变量的多重共线性对参数估计的影响很小，可以进行参数估计。

| Output Statistics | | | | | | | | | | |
|-------------------|--------------------|-----------------|------------------------|----------|--------------------|------------------|-------------|--|----|----------|
| Obs | Dependent Variable | Predicted Value | Std Error Mean Predict | Residual | Std Error Residual | Student Residual | -2 -1 0 1 2 | | | Cook's D |
| 1 | 137.0000 | 132.8153 | 3.9699 | 4.1847 | 23.713 | 0.176 | | | | 0.000 |
| 2 | 162.0000 | 166.4940 | 5.0776 | -4.4940 | 23.501 | -0.191 | | | | 0.001 |
| 3 | 134.0000 | 137.7891 | 3.9682 | -3.7891 | 23.713 | -0.160 | | | | 0.000 |
| 4 | 188.0000 | 180.3985 | 6.3595 | 7.6015 | 23.187 | 0.328 | | | | 0.004 |
| 5 | 138.0000 | 116.5507 | 3.3980 | 21.4493 | 23.802 | 0.901 | | | * | 0.008 |
| 6 | 215.0000 | 204.5898 | 6.5570 | 10.4102 | 23.132 | 0.450 | | | | 0.008 |
| 7 | 171.0000 | 192.6457 | 5.7328 | -21.6457 | 23.350 | -0.927 | | | * | 0.026 |
| 8 | 148.0000 | 154.7867 | 6.6333 | -6.7867 | 23.110 | -0.294 | | | | 0.004 |
| 9 | 197.0000 | 173.3980 | 5.3497 | 23.6020 | 23.440 | 1.007 | | | ** | 0.026 |
| 10 | 113.0000 | 197.3468 | 6.9496 | -84.3468 | 23.017 | -3.665 | ***** | | | 0.612 |

| | | | | | | | | | | |
|----|----------|----------|---------|----------|--------|---------|--|-----|--|-------|
| 11 | 145.0000 | 136.4216 | 3.9791 | 8.5784 | 23.712 | 0.362 | | | | 0.002 |
| 12 | 81.0000 | 82.4932 | 16.4258 | -1.4932 | 17.557 | -0.0850 | | | | 0.003 |
| 13 | 185.0000 | 167.9525 | 4.8196 | 17.0475 | 23.555 | 0.724 | | * | | 0.011 |
| 14 | 157.0000 | 164.2249 | 5.1743 | -7.2249 | 23.480 | -0.308 | | | | 0.002 |
| 15 | 197.0000 | 192.5964 | 9.1670 | 4.4036 | 22.227 | 0.198 | | | | 0.003 |
| 16 | 156.0000 | 157.3949 | 8.0612 | -1.3949 | 22.651 | -0.0616 | | | | 0.000 |
| 17 | 156.0000 | 145.3769 | 4.1943 | 10.6231 | 23.674 | 0.449 | | | | 0.003 |
| 18 | 154.0000 | 143.7252 | 4.2341 | 10.2748 | 23.667 | 0.434 | | | | 0.003 |
| 19 | 144.0000 | 113.3630 | 7.1755 | 30.6370 | 22.947 | 1.335 | | ** | | 0.087 |
| 20 | 90.0000 | 114.4218 | 6.5969 | -24.4218 | 23.120 | -1.056 | | * | | 0.045 |
| 21 | 215.0000 | 170.0095 | 5.4416 | 44.9905 | 23.419 | 1.921 | | *** | | 0.100 |
| 22 | 184.0000 | 177.9002 | 5.1813 | 6.0998 | 23.478 | 0.260 | | | | 0.002 |
| 23 | 118.0000 | 110.2093 | 3.5634 | 7.7907 | 23.778 | 0.328 | | | | 0.001 |
| 24 | 127.0000 | 155.1125 | 4.7275 | -28.1125 | 23.574 | -1.193 | | * | | 0.029 |
| 25 | 137.0000 | 130.9761 | 3.8366 | 6.0239 | 23.735 | 0.254 | | | | 0.001 |
| 26 | 126.0000 | 143.9733 | 4.3186 | -17.9733 | 23.652 | -0.760 | | * | | 0.010 |
| 27 | 130.0000 | 129.4267 | 4.0289 | 0.5733 | 23.703 | 0.0242 | | | | 0.000 |
| 28 | 135.0000 | 143.6892 | 5.1996 | -8.6892 | 23.474 | -0.370 | | | | 0.003 |
| 29 | 188.0000 | 149.8848 | 4.2875 | 38.1152 | 23.658 | 1.611 | | *** | | 0.043 |
| 30 | 122.0000 | 140.1075 | 6.9552 | -18.1075 | 23.015 | -0.787 | | * | | 0.028 |

第六部分输出残差分析的结果,其中第一列为观测号,第二列至第四列分别为因变量的观测值、预测值与预测值的标准误,第五列和第六列为残差及其标准误,第七列为学生化残差,第八列为学生化残差在各个观测点的粗略分布图,第九列为 Cook's D 统计量。当学生化残差大于 2 时,对应的观测点可能就是异常点,本例中第十例观测学生化残差的绝对值为 3.665,很可能是异常点,所以要对该点的情况进行仔细考察,发现该点出现的原因,结合专业知识做出相应的处理。

| | |
|------------------------------|----------|
| Sum of Residuals | 23.92565 |
| Sum of Squared Residuals | 16186 |
| Predicted Residual SS(PRESS) | 18560 |

第七部分输出三个残差统计量,分别是残差和、残差平方和与预测残差平方和(PRESS)。最后一项(PRESS)的含义是:每次去掉一个观测点建立回归方程,并将去掉的那点的全部自变量的取值代入回归方程计算出与该点对应的因变量的预测值,进而求出该点上因变量的残差(即因变量的观测值与刚算出的预测值之差)的平方。这样的工作从第一个观测点到最后一个观测点分别做一次,将全部点上的残差平方相加,就称为预测残差平方和(PRESS)。

这是在图形输出窗口产生的残差图,可以看到图中各点基本上是随机均匀地分布在 $y=0$ 的两侧,没有呈现出某种特别的趋势,其中只有一例残差较大,正是在第六部分结果中学生化残差绝对值为 3.665 的第十例观测。

专业结论:根据现有资料,经多重线性回归分析,得出载脂蛋白 B 和载脂蛋白 C 对低密度脂蛋白中胆固醇含量的影响具有统计学意义。最终模型得出决定系数 $R^2=0.9776>0.5$,说明由这两种载脂蛋白的含量推测低密度脂蛋白中胆固醇含量有较大的实用价值。

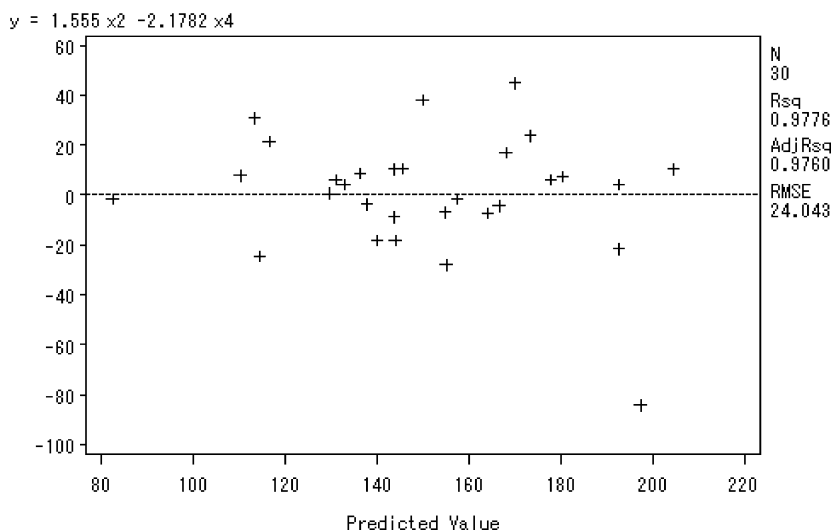


图 19-1 反映多重线性回归方程对资料拟合效果如何的残差图

19.10 REG 过程语法简介

多重线性回归分析主要是用 SAS 软件中的 REG 过程来实现,在该过程中除了对全模型进行计算外(不筛选自变量),还可以运用 8 种方法对自变量进行筛选,并且提供回归系数的假设检验及回归中各种统计量的计算结果。此外,还可以进行共线性诊断及按不同的需求绘制不同的散点图来进行回归方程的选择。

REG 过程的语法如下:

```
PROC REG <options >;
  <label: >MODEL dependents = <regressors > </ options >;
  BY variables ;
  FREQ variable ;
  ID variables ;
  VAR variables ;
  WEIGHT variable ;
  ADD variables ;
  DELETE variables ;
  <label: >MTEST <equation,...,equation> </ options >;
  OUTPUT <OUT = SAS-data-set > keyword = names
  <...keyword = names >;
  PAINT <condition |ALLOBS >
  </ options > |<STATUS |UNDO>;
  PLOT <yvariable* xvariable> <=symbol> <=symbol> </ options >;
  PRINT <options > <ANOVA> <MODELDATA>;
  REFIT;
  RESTRICT equation, ...,equation ;
  REWEIGHT <condition |ALLOBS> </ options > |<STATUS |UNDO>;
  <label: >TEST equation, <,...,equation> </ option >;
```

由以上看出 REG 过程的语句较多,并且每条语句中还有众多的选择项,看起来很复杂。

但是对多重线性回归分析来说,常用的也就有 3~4 条语句。下面介绍常用的 MODEL、OUTPUT、PAINT、PLOT,其余的可参照有关的资料。

“PROC REG <options>;”是过程步的开始,其后有众多的选项,常用的有:

DATA = 说明所要分析的数据集的名称,如“DATA = A”说明要分析数据集 A。

OUTEST = 将估计的回归系数及回归方程中的统计量形成数据集,如“OUTEST = AA”说明将需要的统计量形成数据集 AA。

COVOUT 将估计参数的方差协方差矩阵输入到“OUTEST = 数据集名称”所指定的数据集中。

LINEPRINTER 以文本的形式绘制散点图。

“<label> MODEL dependents = <regressors> </options>;”是 REG 过程的核心部分,是将所要进行拟合的因变量和自变量列出,并在“/”后的选项中列出分析的要求,如自变量的选择方法以及需要的统计量等。例如:

```
Proc reg data = A;
M1: model y = x1 - x6 / selection = forward;
M2: model y = x1 x3 x6 / AIC BIC CP;
Run;
```

MODEL 的选择项非常多,除了进行自变量选择的参数外,还有很多模型及观测的统计量,可根据需要进行选择。

“PAINT <condition | ALLOBS>”是对满足条件的观测进行标记,但是它只有在 REG 中的选项 LINEPRINTER 存在时才起作用,否则将被忽略掉。与 PLOT 联用时,可以在图中突出显示符合条件的点。例如:

```
paint Name = 'Henry' / symbol = 'H';
plot student.* p.;
```

可以将数据集中名为“Henry”所对应的点在残差图中显示为“H”。

“PLOT <yvariable * xvariable> <= symbol> <= symbol> </options>;”是进行散布图的绘制,如果在 REG 过程后有 LINEPRINTER 选项,则以文本形式绘制散布图,否则以图形的形式绘制散布图。在 PLOT 后也有很多的选项,其中最重要的是 <yvariable * xvariable>,它确定散布图纵横坐标的内容,并以“*”分开,如“PLOT r. * p.”则可绘制残差图,其中纵轴表示残差,横轴表示预测值;<= symbol>选项是在有 LINEPRINTER 选项的情况下才起作用,用来确定散布图中点的符号。

此外,使用 REG 过程进行数据处理时,有一个很方便的功能就是可以在不退出 REG 过程的情况下与用户及时地进行交互,这样大大方便了用户对模型进行调整。但是,在该过程中并不是所有的语句都可以进行交互,只有 ADD、DELETE、MODEL、MTEST、OUTPUT、PAINT、PLOT、PRINT、REFIT、RESTRICT、REWEIGHT 和 TEST 语句可以在不退出 REG 过程的情况下进行交互。例如在对全模型计算回归系数后,如果觉得某一自变量 x_i 不合适,想将其从全模型中去除再进行计算,可在不退出 REG 过程的情况下,加入语句“DELETE x_i ; PRINT; run;”后,运行该语句即可。例如,设数据集 A 中有 6 个自变量,用如下语句进行分析:

```
proc reg data = A;
model y = x1 - x5;
run;
```

```
Delete x3;  
print; run;  
Add x6;  
print; run;
```

在运行了以上程序中，前3行先是将自变量 x1 至 x5 选入模型，如果想将自变量 x3 去除后重新拟合模型，则直接在选中第4行运行即可；如果想将自变量 x6 选入模型，则直接运行第5行程序即可，而不必将 reg 过程中的语句重写，此时模型中的自变量为 x1、x2、x4、x5 和 x6。

19.11 本章小结

在应用多重线性回归时，自变量的筛选、共线性诊断、异常点诊断以及残差分析是特别需要引起注意的地方，这几个方面在 SAS 的 REG 过程中都可以通过一定的语句或选项来实现，分析时最好不要将它们忽略。在处理实际问题时，模型的准确选择在很大程度上要依赖于所研究问题本身的专业知识和实践经验，当应用某一种准则和方法选出的一个最优回归方程与实际问题的专业理论不一致时，应重新考虑统计结论，仔细从数据中寻找是否含有异常点、共线性等。

(葛毅 柳伟伟 胡良平)

第 20 章 主成分回归分析

主成分回归分析是将多个彼此相关、信息重叠的指标通过适当的线性组合，使之成为彼此独立而又提取了原指标变异信息并带有特定专业含义的综合潜变量，即主成分，建立潜变量和因变量间的线性回归方程，再将回归方程中的潜变量转换为原自变量的一种多元统计方法。多重线性回归分析和主成分回归分析都是用于分析单组设计多元定量资料的统计学方法，但主成分回归分析的前提条件是自变量之间存在严重的共线性。本章将对主成分回归分析的 SAS 实现方法予以介绍。

20.1 问题、数据及统计分析方法的选择

20.1.1 问题与数据

【例 20-1】 某种水泥在凝固时放出的热量 Y (卡/克) 与该物质中的下列 4 种化学成分有关: X_1 : $3\text{CaO} \cdot \text{Al}_2\text{O}_3$ 的成分(%); X_2 : $3\text{CaO} \cdot \text{SiO}_2$ 的成分(%); X_3 : $4\text{CaO} \cdot \text{Al}_2\text{O}_3 \cdot \text{Fe}_2\text{O}_3$ 的成分(%); X_4 : $2\text{CaO} \cdot \text{SiO}_2$ 的成分(%)。做 Y 对 4 个自变量 X_1 、 X_2 、 X_3 、 X_4 的线性回归方程。数据结构如表 20-1 所示。

表 20-1 某种水泥在凝固时放出的热量 Y 与其 4 种成分的记录

| 样品编号 | X_1 (%) | X_2 (%) | X_3 (%) | X_4 (%) | Y (卡/克) |
|------|-----------|-----------|-----------|-----------|-----------|
| 1 | 7 | 26 | 6 | 60 | 78.5 |
| 2 | 1 | 29 | 15 | 52 | 74.3 |
| 3 | 11 | 56 | 8 | 20 | 104.3 |
| 4 | 11 | 31 | 8 | 47 | 87.6 |
| 5 | 7 | 52 | 6 | 33 | 95.9 |
| 6 | 11 | 55 | 9 | 22 | 109.2 |
| 7 | 3 | 71 | 17 | 6 | 102.7 |
| 8 | 1 | 31 | 22 | 44 | 72.5 |
| 9 | 2 | 54 | 18 | 22 | 93.1 |
| 10 | 21 | 47 | 4 | 26 | 115.9 |
| 11 | 1 | 40 | 23 | 34 | 83.8 |
| 12 | 11 | 66 | 9 | 12 | 113.3 |
| 13 | 10 | 68 | 8 | 12 | 109.4 |

【例 20-2】 为了了解和预测人体吸入氧气的效率，收集了 30 名中年男性的健康状况调查资料。共调查了 7 个指标: 吸氧效率(y), 年龄(x_1 , 岁), 体重(x_2 , kg), 跑 1.5km 所需时间(x_3 , min), 休息时的心率(x_4 , 次/分), 跑步时的心率(x_5 , 次/分), 最高心率(x_6 , 次/分), 数据见表 20-2。该资料中，吸氧的效率是因变量，其余 6 个为自变量，试建立人体吸氧效率模型。

表 20-2 30 名中年男性的健康状况调查资料

| id | y | x ₁ (岁) | x ₂ (kg) | x ₃ (min) | x ₄ (次/分) | x ₅ (次/分) | x ₆ (次/分) |
|-----|--------|--------------------|---------------------|----------------------|----------------------|----------------------|----------------------|
| 1 | 44.609 | 44 | 89.47 | 11.37 | 62 | 178 | 182 |
| 2 | 45.313 | 40 | 75.07 | 10.07 | 62 | 185 | 185 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 30 | 47.920 | 48 | 61.24 | 11.50 | 52 | 170 | 176 |
| 31 | 47.467 | 52 | 82.78 | 10.50 | 53 | 170 | 172 |

备注：id 表示观测对象编号，y 表示吸氧效率，x₁ 表示年龄，x₂ 表示体重，x₃ 表示跑 1.5km 所需时间，x₄ 表示休息时的心率，x₅ 表示跑步时的心率，x₆ 表示最高心率。

20.1.2 对数据结构的分析

例 20-1 数据为 13 份某种水泥在凝固时放出的热量及其 4 种化学成分的质量百分含量测量值。假定此 13 份某种水泥在密度、质量、初始温度等方面具有较好的同质性，则此资料属于单组设计五元定量资料。

例 20-2 数据为 30 名中年男性的健康状况调查资料，包括吸氧效率、年龄、体重、跑 1.5km 所需时间、休息时的心率、跑步时的心率、最高心率共 7 项定量指标。假定 30 名中年男性的年龄相差不大，身体健康状况差不多(比如：均健康良好)，即 30 名中年男性在年龄、身体健康状况等方面具有较好的同质性，则例 20-2 资料为单组设计七元定量资料。

20.1.3 分析目的及统计分析方法的选择

例 20-1 的分析目的是：寻找某种水泥凝固时放出的热量与 4 种化学成分质量百分含量之间的关系。例 20-2 的分析目的是：寻找中年健康男性的吸氧效率与年龄、体重、跑 1.5km 所需时间、休息时的心率、跑步时的心率、最高心率 6 个定量变量之间的关系。两个例题的分析目的表面上不一样，本质上却是完全相同的。有可能实现这种分析目的的分析方法有多重线性回归分析、二次曲面回归分析、病态数据回归分析、岭回归分析、主成分回归分析等。这里，主要介绍主成分回归分析的 SAS 实现方法。感兴趣的读者，也可以采用其他方法对此数据进行分析。至于哪一种方法的分析结果是最佳的，针对具体的问题，需要得到各种方法的分析结果后再做进一步的比较，方可得出结论。

20.2 主成分回归分析应用场合及关键技术

进行多重线性回归分析时有一个重要的前提条件，即全部自变量之间应当互相独立。然而，实际资料往往不满足此前提条件。设法消除自变量之间存在的多重共线性成了进行多重线性回归分析的一项艰巨任务。最常见的解决办法有：(1)对全部定量自变量进行主成分分析，获得互相独立的主成分变量，将其视为“新的自变量”；(2)直接采用“岭回归分析”，通过引入“岭参数”间接消除自变量之间的多重共线性对回归方程造成的影响；(3)采用投影寻踪回归分析，通过引入正交变换的技术，消除自变量之间的多重共线性对回归方程造成的影响。到目前为止，SAS 软件可以采取前两种方法，本章介绍主成分回归分析法。下一章介绍岭回归分析法。

主成分回归分析的关键技术就是对全部定量自变量进行主成分分析，具体做法参见本书第 36 章，此处从略。

20.3 主成分回归分析

20.3.1 对例 20-1 资料进行主成分回归分析

分析时, 先进行多重线性回归分析, 并做共线性诊断; 若存在共线性, 则采用主成分回归分析。SAS 程序如下: 名为 sastjfx20_1.sas。

```

data ex20_1;                                /* 1 */
  input x1 - x4 y @@ ;
  cards;
    7 26 6 60 78.5
    1 29 15 52 74.3
    11 56 8 20 104.3
    11 31 8 47 87.6
    7 52 6 33 95.9
    11 55 9 22 109.2
    3 71 17 6 102.7
    1 31 22 44 72.5
    2 54 18 22 93.1
    21 47 4 26 115.9
    1 40 23 34 83.8
    11 66 9 12 113.3
    10 68 8 12 109.4
;
run;
ods html;
proc reg data = ex20_1;                      /* 2 */
  model y = x1 - x4 / collin collinoint;
  title 'regression analysis';
run;
proc princomp data = ex20_1 out = pc2
  prefix = z;                                /* 3 */
  var x1 - x4;
  title 'principal component analysis';
run;
proc print data = pc2;                       /* 4 */
  title 'principal component scores';
run;
proc reg data = pc2;                         /* 5 */
  model y = z1 z2 / stb;
  title 'regression analysis on principal components';
run;
ods html close;

```

程序第 1 步建立数据集, 命名为 ex20_1。以 x1 - x4 分别表示 4 种化学成分, y 表示水泥凝固时放出的热量。第 2 步对资料进行多重线性回归分析, 并使用 collin 和 collinoint 选项对自变量进行共线性诊断。第 3 步对 x1 - x4 进行主成分分析, 并将原始数据及主成分得分一起输出到数据集 pc2 中。第 4 步将数据集 pc2 中的内容输出到 output 窗口。第 5 步以因变量 y 对 x1 - x4 的第一、二主成分做多重线性回归分析, 并输出回归方程的标准化回归系数(stb 选项)。

SAS 输出结果:

| The REG Procedure | | | | | |
|-----------------------|----|----------------|-------------|---------|--------|
| Model: MODEL1 | | | | | |
| Dependent Variable: y | | | | | |
| Analysis of Variance | | | | | |
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 4 | 2667.89944 | 666.97486 | 111.48 | <.0001 |
| Error | 8 | 47.86364 | 5.98295 | | |
| Corrected Total | 12 | 2715.76308 | | | |
| Root MSE | | 2.44601 | R-Square | 0.9824 | |
| Dependent Mean | | 95.42308 | Adj R-Sq | 0.9736 | |
| Coeff Var | | 2.56333 | | | |

方差分析的结果显示，模型总体拟合数据较好 ($F = 111.48$, $P < 0.001$, $R^2 = 0.9824$)。

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 62.40537 | 70.07096 | 0.89 | 0.3991 |
| x1 | 1 | 1.55110 | 0.74477 | 2.08 | 0.0708 |
| x2 | 1 | 0.51017 | 0.72379 | 0.70 | 0.5009 |
| x3 | 1 | 0.10191 | 0.75471 | 0.14 | 0.8959 |
| x4 | 1 | -0.14406 | 0.70905 | -0.20 | 0.8441 |

多重线性回归分析参数估计的结果。截距项及各变量的系数与 0 的差异都没有统计学意义 ($P > 0.05$)。这与对总模型评价的方差分析结果相违背，说明自变量间可能存在共线性关系。

| Collinearity Diagnostics | | | | | | | |
|--------------------------|------------|-----------|-------------------------|------------|------------|------------|------------|
| Number | Eigenvalue | Condition | Proportion of Variation | | | | |
| | | Index | Intercept | x1 | x2 | x3 | x4 |
| 1 | 4.11970 | 1.00000 | 0.00000551 | 0.00036889 | 0.00001833 | 0.00021022 | 0.00003641 |
| 2 | 0.55389 | 2.72721 | 8.812348E-8 | 0.01004 | 0.00001265 | 0.00266 | 0.00010070 |
| 3 | 0.28870 | 3.77753 | 3.060952E-7 | 0.00057551 | 0.00031981 | 0.00159 | 0.00168 |
| 4 | 0.03764 | 10.46207 | 0.00012679 | 0.05745 | 0.00278 | 0.04569 | 0.00088373 |
| 5 | 0.00006614 | 249.57825 | 0.99987 | 0.93157 | 0.99687 | 0.94985 | 0.99730 |

| Collinearity Diagnostics (intercept adjusted) | | | | | | | |
|---|------------|-----------|-------------------------|------------|---------|------------|--|
| Number | Eigenvalue | Condition | Proportion of Variation | | | | |
| | | Index | x1 | x2 | x3 | x4 | |
| 1 | 2.23570 | 1.00000 | 0.00263 | 0.00055897 | 0.00148 | 0.00047533 | |
| 2 | 1.57607 | 1.19102 | 0.00427 | 0.00042729 | 0.00495 | 0.00045729 | |
| 3 | 0.18661 | 3.46134 | 0.06352 | 0.00208 | 0.04650 | 0.00072440 | |
| 4 | 0.00162 | 37.10634 | 0.92958 | 0.99693 | 0.94707 | 0.99834 | |

共线性诊断的结果。因为截距项与 0 的差异无统计学意义，所以应查看第一部分的结果，即未对截距项校正的共线性诊断。最大条件指数为 $249.57825 \gg 10$ ，且所有自变量的方差分量均接近于 1，说明自变量间存在严重的共线性。

| principal component analysis | | | | |
|------------------------------|-------------|-------------|-------------|-------------|
| The PRINCOMP Procedure | | | | |
| Simple Statistics | | | | |
| | x1 | x2 | x3 | x4 |
| Mean | 7.461538462 | 48.15384615 | 11.76923077 | 30.00000000 |
| StD | 5.882394419 | 15.56088126 | 6.40512615 | 16.73817991 |

自变量的均数和标准差。据此，可写出 $x_1 - x_4$ 的标准化变量，记为 $X_1 - X_4$ ，则有：

$$X_1 = \frac{x_1 - 7.461538}{5.882394}, \quad X_2 = \frac{x_2 - 48.153846}{15.560881}$$
$$X_3 = \frac{x_3 - 11.769231}{6.405126}, \quad X_4 = \frac{x_4 - 30.0}{16.73818}$$

Correlation Matrix

| | x1 | x2 | x3 | x4 |
|----|--------|--------|--------|--------|
| x1 | 1.0000 | 0.2286 | -.8241 | -.2454 |
| x2 | 0.2286 | 1.0000 | -.1392 | -.9730 |
| x3 | -.8241 | -.1392 | 1.0000 | 0.0295 |
| x4 | -.2454 | -.9730 | 0.0295 | 1.0000 |

自变量间的相关系数矩阵。

Eigenvalues of the Correlation Matrix

| | Eigenvalue | Difference | Proportion | Cumulative |
|---|------------|------------|------------|------------|
| 1 | 2.23570403 | 0.65963796 | 0.5589 | 0.5589 |
| 2 | 1.57606607 | 1.38945992 | 0.3940 | 0.9529 |
| 3 | 0.18660615 | 0.18498240 | 0.0467 | 0.9996 |
| 4 | 0.00162375 | | 0.0004 | 1.0000 |

相关系数矩阵的特征值、贡献率和累积贡献率。从主成分累积贡献率来看,前两个主成分包含了原四个自变量信息的 95.29%。因此,可选择前两个主成分替代原来的四个自变量。

Eigenvectors

| | z1 | z2 | z3 | z4 |
|----|----------|----------|----------|----------|
| x1 | 0.475955 | -.508979 | 0.675500 | 0.241052 |
| x2 | 0.563870 | 0.413931 | -.314420 | 0.641756 |
| x3 | -.394067 | 0.604969 | 0.637691 | 0.268466 |
| x4 | -.547931 | -.451235 | -.195421 | 0.676734 |

特征向量。据此,可以写出由标准化变量所表达的各主成分的关系式,即

$$Z_1 = 0.475955X_1 + 0.56387X_2 - 0.394067X_3 - 0.547931X_4$$

$$Z_2 = -0.508979X_1 + 0.413931X_2 + 0.604969X_3 - 0.451235X_4$$

$$Z_3 = 0.6755X_1 - 0.31442X_2 + 0.637691X_3 - 0.195421X_4$$

$$Z_4 = 0.241052X_1 + 0.641756X_2 + 0.268466X_3 + 0.676734X_4$$

式中, X_i 为原自变量的标准化变量, $i = 1, 2, 3, 4$ 。

principal component scores

| Obs | x1 | x2 | x3 | x4 | y | z1 | z2 | z3 | z4 |
|-----|----|----|----|----|-------|----------|----------|----------|-----------|
| 1 | 7 | 26 | 6 | 60 | 78.5 | -1.46724 | -1.90304 | -0.53000 | 0.038530 |
| 2 | 1 | 29 | 15 | 52 | 74.3 | -2.13583 | -0.23835 | -0.29019 | -0.029833 |
| 3 | 11 | 56 | 8 | 20 | 104.3 | 1.12987 | -0.18388 | -0.01071 | -0.093701 |
| 4 | 11 | 31 | 8 | 47 | 87.6 | -0.65990 | -1.57677 | 0.17920 | -0.033116 |
| 5 | 7 | 52 | 6 | 33 | 95.9 | 0.35876 | -0.48354 | -0.74012 | 0.019187 |
| 6 | 11 | 55 | 9 | 22 | 109.2 | 0.96664 | -0.16994 | 0.08570 | -0.012167 |
| 7 | 3 | 71 | 17 | 6 | 102.7 | 0.93071 | 2.13482 | -0.17299 | 0.008295 |
| 8 | 1 | 31 | 22 | 44 | 72.5 | -2.23214 | 0.69167 | 0.45972 | 0.022606 |
| 9 | 2 | 54 | 18 | 22 | 93.1 | -0.35152 | 1.43225 | -0.03156 | -0.044988 |
| 10 | 21 | 47 | 4 | 26 | 115.9 | 1.66254 | -1.82810 | 0.85119 | 0.019837 |
| 11 | 1 | 40 | 23 | 34 | 83.8 | -1.64018 | 1.29511 | 0.49418 | 0.031389 |
| 12 | 11 | 66 | 9 | 12 | 113.3 | 1.69259 | 0.39225 | -0.01981 | 0.037185 |
| 13 | 10 | 68 | 8 | 12 | 109.4 | 1.74568 | 0.43753 | -0.27462 | 0.036776 |

以上是原始自变量、因变量和主成分得分。

| | | | | | |
|---|----|----------------|-------------|---------|--------|
| regression analysis on principal components | | | | | |
| The REG Procedure | | | | | |
| Model: MODEL1 | | | | | |
| Dependent Variable: y | | | | | |
| Analysis of Variance | | | | | |
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 2 | 2620.77920 | 1310.38960 | 137.96 | <.0001 |
| Error | 10 | 94.98387 | 9.49839 | | |
| Corrected Total | 12 | 2715.76308 | | | |
| Root MSE | | 3.08195 | R-Square | 0.9650 | |
| Dependent Mean | | 95.42308 | Adj R-Sq | 0.9580 | |
| Coeff Var | | 3.22977 | | | |

对因变量 y 和主成分 Z_1 、 Z_2 进行多重线性回归分析的结果。模型总体较好地拟合了数据 ($F = 137.96$, $P < 0.0001$, $R^2 = 0.9650$)。

| | | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|-----------------------|
| Parameter Estimates | | | | | | |
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t | Standardized Estimate |
| Intercept | 1 | 95.42308 | 0.85478 | 111.63 | <.0001 | 0 |
| z1 | 1 | 9.88309 | 0.59501 | 16.61 | <.0001 | 0.98230 |
| z2 | 1 | 0.12499 | 0.70867 | 0.18 | 0.8635 | 0.01043 |

多重线性回归分析参数估计的结果。截距项及 Z_1 的系数与 0 的差异有统计学意义 ($P < 0.05$)，但 Z_2 的系数与 0 的差异无统计学意义 ($P = 0.8635 > 0.05$)。本模型中，最后只保留了一个主成分 Z_1 ，它包含了原四个自变量 55.89% 的信息，拟合效果不是太好。因为这是以回归为目的，应进行变量筛选。在因变量与 $Z_1 - Z_4$ 做多重线性回归分析时，运用 stepwise (sle = 0.30、sls = 0.05)、forward (sle = 0.05) 和 backward (sls = 0.05) 方法进行自变量的筛选 (需修改程序 sastjfx20_1.sas 的第 5 步，重新运行)，最终得到的结果如下 (对本资料，三种方法结果一致)。

| | | | | | |
|---|----|----------------|-------------|---------|--------|
| regression analysis on principal components | | | | | |
| The REG Procedure | | | | | |
| Model: MODEL1 | | | | | |
| Dependent Variable: y | | | | | |
| Analysis of Variance | | | | | |
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 2 | 2666.93998 | 1333.46999 | 273.12 | <.0001 |
| Error | 10 | 48.82310 | 4.88231 | | |
| Corrected Total | 12 | 2715.76308 | | | |
| Root MSE | | 2.20959 | R-Square | 0.9820 | |
| Dependent Mean | | 95.42308 | Adj R-Sq | 0.9784 | |
| Coeff Var | | 2.31558 | | | |

| Parameter Estimates | | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|-----------------------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t | Standardized Estimate |
| Intercept | 1 | 95.42308 | 0.61283 | 155.71 | <.0001 | 0 |
| z1 | 1 | 9.88309 | 0.42659 | 23.17 | <.0001 | 0.98230 |
| z3 | 1 | 4.55479 | 1.47659 | 3.08 | 0.0115 | 0.13079 |

模型总体较好地拟合了数据 ($F = 273.12$, $P < 0.0001$, $R^2 = 0.9820$)。截距项及 Z_1 、 Z_3 的系数与 0 的差异均有统计学意义。因此, 回归方程为:

$$Y = 95.42308 + 9.88309Z_1 + 4.55479Z_3$$

主成分对原自变量信息的涵盖量提高到了 60.56%, 较上述回归模型要好。将 Z_1 、 Z_3 的表达式回代到此回归方程, 得:

$$\begin{aligned} Y &= 95.42308 + 9.88309Z_1 + 4.55479Z_3 \\ &= 95.42308 + 7.780667X_1 + 4.140661X_2 - 0.99005X_3 - 6.30535X_4 \end{aligned}$$

将标准化变量还原为原始自变量, 可得到因变量对原始变量的回归模型:

$$Y = 85.8605 + 1.32270x_1 + 0.26609x_2 - 0.15457x_3 - 0.3767x_4$$

20.3.2 对例 20-2 数据进行分析

分析时, 先进行多重线性回归分析, 并做共线性诊断; 若自变量间确实存在共线性, 则采用主成分回归分析。SAS 程序如下, 名为 sastjfx20_2.sas。

```
data sastjfx20_2;          /* 1 */
  infile 'd:\sastjfx\ex20_2.txt';
  input y x1 -x6;
run;
ods html;
proc reg;                  /* 2 */
  model y=x1-x6/collin collinoint
  selection=stepwise;
run;
proc princomp out=pc2 prefix=z;
  var x1-x6;               /* 3 */
run;

proc print data=pc2;       /* 4 */
  title 'principal component
  scores';
run;
proc reg data=pc2;         /* 5 */
  model y =z1-z4/selection
  =stepwise
  sle=0.30 sls=0.05;
  title 'Regression analysis on
  principal components';
run;
ods html close;
```

将原始数据以 31 行 7 列形式输入计算机, 以文本文件格式存入 D 盘上的 sastjfx 文件夹内, 数据文件名为 ex20_2.txt。第 1 步通过 infile 语句读入存储在 D 盘 sastjfx 文件夹下的外部数据文件。第 2 步对资料进行多重线性回归分析, 采用逐步回归法进行自变量的筛选 (sls 和 sle 采用 SAS 软件默认的 0.15 检验水准), 并对自变量间是否存在共线性问题进行诊断。第 3 步对自变量进行主成分分析, 并将主成分得分输出到数据集 pc2 中。第 4 步输出数据集 pc2 的内容至 output 窗口。第 5 步对资料进行主成分回归分析, 分析的数据集为 pc2, 建立因变量 y 与前四个主成分 $z_1 - z_4$ 间的多重线性回归方程, 采用逐步回归法主成分的筛选 (sle = 0.30、sls = 0.05)。

SAS 输出结果:

```
The REG Procedure
Model: MODEL1
Dependent Variable: y
```

| Analysis of Variance | | | | | |
|----------------------|----|----------------|-------------|---------|--------|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 4 | 717.69550 | 179.42388 | 34.90 | <.0001 |
| Error | 26 | 133.68604 | 5.14177 | | |
| Corrected Total | 30 | 851.38154 | | | |

逐步回归分析筛选变量的最终结果。方差分析的结果显示，模型总体拟合数据较好 ($F = 34.90$, $P < 0.001$)，模型拟合因变量的残差平方和为 133.68604，残差均方为 5.14177。

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|-----------|--------------------|----------------|------------|---------|--------|
| Intercept | 100.07910 | 11.57739 | 384.21858 | 74.72 | <.0001 |
| x1 | -0.21266 | 0.09099 | 28.08629 | 5.46 | 0.0274 |
| x3 | -2.76824 | 0.33138 | 358.80967 | 69.78 | <.0001 |
| x5 | -0.33957 | 0.11555 | 44.40268 | 8.64 | 0.0068 |
| x6 | 0.25535 | 0.13188 | 19.27645 | 3.75 | 0.0638 |

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

采用逐步回归法进行多重线性回归分析参数估计的结果。截距项及自变量 x_1 、 x_3 、 x_5 的系数与 0 的差异都有统计学意义 ($P < 0.05$)，但自变量 x_6 的系数与 0 的差异没有统计学意义 ($P = 0.0638 > 0.05$)。

| Summary of Stepwise Selection | | | | | | | | |
|-------------------------------|------------------|------------------|----------------|------------------|----------------|---------|---------|--------|
| Step | Variable Entered | Variable Removed | Number Vars In | Partial R-Square | Model R-Square | C(p) | F Value | Pr > F |
| 1 | x3 | | 1 | 0.7434 | 0.7434 | 15.5244 | 84.01 | <.0001 |
| 2 | x1 | | 2 | 0.0268 | 0.7702 | 13.0817 | 3.27 | 0.0815 |
| 3 | x5 | | 3 | 0.0501 | 0.8203 | 6.7720 | 7.54 | 0.0106 |
| 4 | x6 | | 4 | 0.0226 | 0.8430 | 5.0201 | 3.75 | 0.0638 |

逐步回归分析中变量筛选的过程。 x_3 、 x_1 、 x_5 、 x_6 依次进入回归模型。

| Collinearity Diagnostics | | | | | | | |
|--------------------------|------------|-----------|-------------------------|------------|------------|------------|------------|
| Number | Eigenvalue | Condition | Proportion of Variation | | | | |
| | | Index | x1 | x3 | x5 | x6 | |
| 1 | 4.97401 | 1.00000 | 0.00004993 | 0.00034656 | 0.00053099 | 0.00001734 | 0.00001270 |
| 2 | 0.01331 | 19.32993 | 0.00002110 | 0.48063 | 0.02310 | 0.00399 | 0.00270 |
| 3 | 0.01135 | 20.93400 | 0.00912 | 0.00277 | 0.90263 | 0.00217 | 0.00230 |
| 4 | 0.00115 | 65.83120 | 0.72976 | 0.42756 | 0.05225 | 0.08730 | 0.00796 |
| 5 | 0.00018301 | 164.86211 | 0.26105 | 0.08869 | 0.02149 | 0.90652 | 0.98703 |

| Collinearity Diagnostics(intercept adjusted) | | | | | | | |
|--|------------|-----------|-------------------------|---------|---------|------------|--|
| Number | Eigenvalue | Condition | Proportion of Variation | | | | |
| | | Index | x1 | x3 | x5 | x6 | |
| 1 | 2.25188 | 1.00000 | 0.04344 | 0.01833 | 0.02177 | 0.02144 | |
| 2 | 1.16684 | 1.38921 | 0.23497 | 0.42235 | 0.00106 | 0.00003238 | |
| 3 | 0.51921 | 2.08258 | 0.65922 | 0.52940 | 0.02452 | 0.01812 | |
| 4 | 0.06208 | 6.02289 | 0.06237 | 0.02992 | 0.95266 | 0.96041 | |

共线性诊断的结果。因为截距项与 0 的差异有统计学意义, 所以应查看第二部分的结果, 即对截距项校正了的共线性诊断结果。最大条件指数为 $6.02289 < 10$, 自变量 x_5 、 x_6 的方差分量均接近于 1, 说明自变量间存在一定程度的共线性, 但并不严重。所以, 选用多重线性回归分析即可。不过, 需对程序 `sastjfx20_2.sas` 中的第 2 步略做改动, 使变量进入和剔除出模型的标准再严格一些。目前, 程序采用的是 SAS 默认的进入和剔除标准, 即“`sle = 0.15, sls = 0.15`”, 为保证最终模型中的所有变量均有统计学意义, 此标准可改为“`sle = 0.15, sls = 0.05`” (将此语句添加到程序 `sastjfx20_2.sas` 中的第 2 步“`selection = stepwise`”选项后, 并重新运行程序)。这样最终保留在模型中的变量仅有 X_3 , 回归模型为 $y = 82.42177 - 3.31056 \times X_3$, 此模型拟合因变量的残差平方和为 218.48144, 残差均方为 7.53384。

另外, 对本资料也可考虑进行主成分回归分析, 结果如下。

Simple Statistics

| | x1 | x2 | x3 | x4 | x5 | x6 |
|------|-------------|-------------|-------------|-------------|-------------|-------------|
| Mean | 47.80645161 | 77.44451613 | 10.58612903 | 53.45161290 | 169.6451613 | 173.7741935 |
| StD | 5.36916725 | 8.32856764 | 1.38741409 | 7.61944315 | 10.2519864 | 9.1640954 |

这是 6 个自变量的均数和标准差。据此, 可写出 $x_1 - x_6$ 的标准化变量, 记为 $X_1 - X_6$, 则有:

$$X_1 = \frac{x_1 - 47.80645161}{5.36916725}, \quad X_2 = \frac{x_2 - 77.44451613}{8.32856764}, \quad X_3 = \frac{x_3 - 10.58612903}{1.38741409},$$

$$X_4 = \frac{x_4 - 53.45161290}{7.61944315}, \quad X_5 = \frac{x_5 - 169.6451613}{10.2519864}, \quad X_6 = \frac{x_6 - 173.7741935}{9.1640954}$$

Correlation Matrix

| | x1 | x2 | x3 | x4 | x5 | x6 |
|----|--------|--------|--------|--------|--------|--------|
| x1 | 1.0000 | -.2300 | 0.1661 | -.1770 | -.3416 | -.4413 |
| x2 | -.2300 | 1.0000 | 0.1435 | 0.0440 | 0.1815 | 0.2494 |
| x3 | 0.1661 | 0.1435 | 1.0000 | 0.4504 | 0.3136 | 0.2261 |
| x4 | -.1770 | 0.0440 | 0.4504 | 1.0000 | 0.3525 | 0.3051 |
| x5 | -.3416 | 0.1815 | 0.3136 | 0.3525 | 1.0000 | 0.9298 |
| x6 | -.4413 | 0.2494 | 0.2261 | 0.3051 | 0.9298 | 1.0000 |

自变量间的相关系数矩阵。

Eigenvalues of the Correlation Matrix

| | Eigenvalue | Difference | Proportion | Cumulative |
|---|------------|------------|------------|------------|
| 1 | 2.58574195 | 1.27999047 | 0.4310 | 0.4310 |
| 2 | 1.30575148 | 0.38038332 | 0.2176 | 0.6486 |
| 3 | 0.92536816 | 0.17908709 | 0.1542 | 0.8028 |
| 4 | 0.74628107 | 0.36914392 | 0.1244 | 0.9272 |
| 5 | 0.37713715 | 0.31741696 | 0.0629 | 0.9900 |
| 6 | 0.05972018 | | 0.0100 | 1.0000 |

相关系数矩阵的特征值、贡献率和累积贡献率。从主成分累积贡献率来看, 前四个主成分包含了原四个自变量信息的 92.72%。因此, 可选择前四个主成分替代原来的六个自变量。

Eigenvectors

| z1 | z2 | z3 | z4 | z5 | z6 |
|----|----|----|----|----|----|
|----|----|----|----|----|----|

| | | | | | | |
|----|----------|----------|----------|----------|----------|----------|
| x1 | -.322823 | 0.548459 | 0.081492 | 0.548710 | 0.531761 | 0.066973 |
| x2 | 0.235264 | -.195787 | 0.921468 | -.023350 | 0.233486 | -.046411 |
| x3 | 0.280301 | 0.670269 | 0.236942 | 0.073735 | -.639200 | 0.044893 |
| x4 | 0.355846 | 0.427835 | -.136982 | -.664691 | 0.478800 | 0.022458 |
| x5 | 0.562235 | -.044226 | -.208206 | 0.376705 | 0.117649 | -.694866 |
| x6 | 0.564919 | -.163100 | -.161246 | 0.330476 | 0.105108 | 0.712743 |

特征向量。据此，可以写出由标准化变量所表达的各主成分的关系式，即

$$Z_1 = -0.322823X_1 + 0.235264X_2 + 0.280301X_3 + 0.355846X_4 + 0.562235X_5 + 0.564919X_6$$
$$Z_2 = 0.548459X_1 - 0.195787X_2 + 0.670269X_3 + 0.427835X_4 - 0.044226X_5 - 0.1631X_6$$
$$Z_3 = 0.081492X_1 + 0.921468X_2 + 0.236942X_3 - 0.136982X_4 - 0.208206X_5 - 0.161246X_6$$
$$Z_4 = 0.54871X_1 - 0.023350X_2 + 0.073735X_3 - 0.664691X_4 + 0.376705X_5 + 0.330476X_6$$

式中， X_i 为原自变量的标准化变量， $i=1、2、3、4、5、6$ 。

| principal component scores | | | | | | | | | | | | | |
|----------------------------|--------|-----|-------|-------|-----|-----|-----|----------|----------|----------|----------|----------|----------|
| Obs | y | x1 | x2 | x3 | x4 | x5 | x6 | z1 | z2 | z3 | z4 | z5 | z6 |
| 1 | 44.609 | 44 | 89.47 | 11.37 | 62 | 178 | 182 | 2.09142 | 0.00472 | 0.93849 | -0.52316 | 0.32639 | 0.00956 |
| 2 | 45.313 | 40 | 75.07 | 10.07 | 62 | 185 | 185 | 2.23134 | -0.77699 | -1.13239 | -0.59526 | 0.24021 | -0.24328 |
| 3 | 54.297 | 44 | 85.84 | 8.65 | 45 | 156 | 168 | -1.42412 | -1.83448 | 1.07111 | -0.48777 | -0.00353 | 0.29394 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 30 | 47.920 | 48 | 61.24 | 11.50 | 52 | 170 | 176 | -0.19587 | 0.71955 | -1.65412 | 0.33372 | -0.91776 | 0.26707 |
| 31 | 47.467 | 52 | 82.78 | 10.50 | 53 | 170 | 172 | -0.22982 | 0.26602 | 0.67138 | 0.39748 | 0.55993 | -0.14358 |

原始自变量、因变量和主成分得分。

| Analysis of Variance | | | | | |
|----------------------|----|----------------|-------------|---------|--------|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 4 | 678.65180 | 169.66295 | 25.54 | <.0001 |
| Error | 26 | 172.72975 | 6.64345 | | |
| Corrected Total | 30 | 851.38154 | | | |

对因变量 y 和主成分 $Z_1 - Z_4$ 进行多重线性回归分析的结果。模型总体较好地拟合了数据 ($F=25.24, P<0.0001$)。

| Variable | Parameter Estimate | Standard Error | Type II SS | F Value | Pr > F |
|-----------|--------------------|----------------|------------|---------|--------|
| Intercept | 47.37581 | 0.46293 | 69578 | 10473.2 | <.0001 |
| z1 | -1.40346 | 0.29265 | 152.79425 | 23.00 | <.0001 |
| z2 | -3.37719 | 0.41182 | 446.77812 | 67.25 | <.0001 |
| z3 | -1.17067 | 0.48919 | 38.04547 | 5.73 | 0.0242 |
| z4 | -1.35382 | 0.54473 | 41.03395 | 6.18 | 0.0197 |

多重线性回归分析参数估计的结果。截距项及 4 个主成分的系数与 0 的差异都有统计学意义 ($P<0.05$)。因此，回归方程为：

$$Y = 47.37581 - 1.40346Z_1 - 3.37719Z_2 - 1.17067Z_3 - 1.35382Z_4$$

将 $Z_1 - Z_4$ 的表达式回代到此回归方程，得：

$$Y = 47.37581 - 2.23744X_1 - 0.7161X_2 - 3.03422X_3 - 0.88406X_4 - 0.90596X_5 - 0.50066X_6$$

将标准化变量还原为原始自变量，可得到因变量对原始变量的回归模型：

$$Y = 127.795 - 0.41672x_1 - 0.08598x_2 - 2.18696x_3 - 0.11603x_4 - 0.08837x_5 - 0.05463x_6$$

此模型拟合因变量的残差平方和为 172.72975，残差均方为 6.64345，可见其并未比模型 $y = 82.42177 - 3.31056 \times X_3$ 的拟合效果好很多。因此，在自变量不存在严重共线性的情况下，一般不需要进行主成分回归分析。

20.4 本章小结

本章通过实例介绍了用 SAS 软件实现主成分回归分析的方法，并对 SAS 软件给出的输出结果做了较详细的解释。做主成分回归分析的前提条件是自变量之间存在严重的共线性。做主成分回归分析的基本步骤如下：

- (1) 进行多重线性回归分析，并进行共线性诊断。
- (2) 如果自变量之间存在共线性，则可选择进行主成分回归分析，以解决由于共线性的影响，造成回归结果不合理或无法解释的问题。
- (3) 用主成分分析求自变量的主成分和主成分得分，将贡献率小的主成分舍弃。
- (4) 将因变量对保留的主成分得分进行回归分析。
- (5) 将主成分的表达式回代，最后得到因变量与原始自变量的回归模型，并给予专业解释。

(郭东星 周诗国)

第 21 章 用 SAS 实现岭回归分析

在第 19 章中介绍了用 REG 过程进行多重线性回归分析,在进行参数估计时常用最小二乘法。该方法在数据满足 GM(Gauss-Markov)定理时,保证了在线性无偏估计类中的方差最小性。如果进一步假设误差服从正态分布,那么最小二乘法还具有更多更好的性质。但是在实际应用中,许多应用实践表明,有些情况在运用最小二乘法时并不是很理想,在个别情况下可能很不好。自 20 世纪 50 年代特别是 60 年代以来,许多统计学家做了很多努力,试图改进最小二乘估计。Stenin 于 1955 年证明了:当维数大于 2 时,能够找到另外一个估计在某种意义下一致优于最小二乘估计。据此,在后来的发展中统计学家提出了许多新的估计方法,主要有岭估计、Stein 估计、主成分估计以及特征根估计等。这些估计有一个共同的特点就是有偏性,本章主要介绍如何用 SAS 实现回归方程中参数的岭回归估计。

21.1 问题、数据及统计分析方法的选择

21.1.1 问题与数据

【例 21-1】 有人在某地抽样调查了 29 例儿童的血红蛋白与 4 种微量元素的含量。试问:可否用 4 种微量元素(单位都是 $\mu\text{mol/L}$)钙(X_1)、镁(X_2)、铁(X_3)、铜(X_4)来较好地预测血红蛋白($Y, \text{g/L}$)的含量? 数据见表 21-1。

表 21-1 29 例儿童的血红蛋白与 4 种微量元素的含量

| id | y | X_1 | X_2 | X_3 | X_4 | id | y | X_1 | X_2 | X_3 | X_4 |
|----|-------|-------|-------|-------|-------|----|-------|-------|-------|-------|-------|
| 1 | 135.0 | 13.70 | 12.68 | 80.32 | 0.16 | 16 | 85.0 | 13.39 | 11.83 | 52.41 | 0.21 |
| 2 | 102.5 | 17.48 | 15.13 | 73.35 | 0.19 | 17 | 120.0 | 15.06 | 15.70 | 70.60 | 0.18 |
| 3 | 130.0 | 18.09 | 17.51 | 83.65 | 0.26 | 18 | 82.5 | 12.53 | 11.99 | 52.38 | 0.16 |
| 4 | 100.0 | 15.73 | 14.41 | 68.75 | 0.13 | 19 | 117.5 | 13.48 | 14.07 | 72.60 | 0.20 |
| 5 | 137.5 | 13.43 | 21.73 | 76.18 | 0.19 | 20 | 80.0 | 16.30 | 12.33 | 55.99 | 0.16 |
| 6 | 97.5 | 12.16 | 12.55 | 61.38 | 0.15 | 21 | 115.0 | 15.28 | 15.35 | 79.83 | 0.22 |
| 7 | 140.0 | 16.15 | 16.10 | 84.09 | 0.19 | 22 | 78.0 | 14.07 | 12.04 | 50.66 | 0.21 |
| 8 | 95.0 | 13.04 | 11.15 | 58.41 | 0.13 | 23 | 112.5 | 15.01 | 13.84 | 68.59 | 0.14 |
| 9 | 142.5 | 14.67 | 15.48 | 81.72 | 0.16 | 24 | 75.0 | 16.50 | 13.12 | 61.61 | 0.11 |
| 10 | 92.5 | 13.03 | 14.87 | 69.55 | 0.16 | 25 | 110.0 | 17.39 | 16.44 | 74.59 | 0.21 |
| 11 | 127.5 | 10.90 | 10.76 | 70.84 | 0.09 | 26 | 72.5 | 18.44 | 13.54 | 55.94 | 0.18 |
| 12 | 90.0 | 12.40 | 10.45 | 59.27 | 0.14 | 27 | 107.5 | 18.03 | 16.49 | 77.11 | 0.19 |
| 13 | 125.0 | 13.70 | 12.68 | 80.32 | 0.16 | 28 | 70.0 | 11.80 | 11.73 | 52.75 | 0.13 |
| 14 | 87.5 | 15.22 | 12.03 | 46.35 | 0.19 | 29 | 105.0 | 13.75 | 13.57 | 79.80 | 0.14 |
| 15 | 122.5 | 21.49 | 18.00 | 78.78 | 0.28 | | | | | | |

21.1.2 对数据结构的分析

表 21-1 中的数据为 29 例儿童的血红蛋白与 4 种微量元素的含量,只涉及一个组,故从实

验设计的类型来看,属于单组设计。受试对象为 29 例儿童,指标变量有 5 个,分别为血红蛋白含量、钙含量、镁含量、铁含量、铜含量。因此,例 21-1 资料属于单组设计五元定量资料。

21.1.3 分析目的与统计分析方法的选择

例 21-1 的分析目的十分明显,就是想建立一个回归方程,通过钙、镁、铁、铜的含量来比较准确地预测血红蛋白的含量。

要达到这个目的,根据现有数据的特征——单组设计多元定量资料,可能的分析方法有多重线性回归分析、响应曲面回归分析、岭回归分析、病态数据回归分析等。那么究竟哪一种方法最好或者比较好呢?正确的分析方案是:分别用不同的回归分析方法对此资料进行分析,然后对不同回归分析方法得出的结果进行比较,看哪一种方法得到的结果最符合专业实际,也就是最好或者比较好。当然,也有可能现有的各种方法所得到的结果都不具有推广应用的价值,如果这样,则需要寻找别的分析方法。这里,仅介绍采用岭回归分析处理此资料的 SAS 实现方法及结果。

21.2 岭回归分析应用场合及关键技术

岭回归分析一般是在多重线性回归分析的基础上进行的,即先进行多重线性回归分析,在进行多重线性回归分析的过程中,通过共线性诊断判断数据是否存在多重共线性。如果自变量之间存在多重共线性关系,在本书第 20 章已做了介绍,常有三种消除或减弱自变量之间存在的多重共线性关系,本章介绍用岭回归分析来进行参数估计。其关键技术就是在对回归系数进行参数估计时引入一个“岭参数”,变动岭参数的数值,使构建的多重线性回归方程最为合适。

从第 19 章可以得知,多重线性回归方程的回归系数可以表示为

$$\beta = (X'X)^{-1}X'Y \quad (21-1)$$

其中 X 为自变量的 $n \times m$ 阶矩阵, X' 为 X 的转置, $(X'X)$ 为对称的 $m \times m$ 方阵, $(X'X)^{-1}$ 为 $X'X$ 的逆矩阵, Y 为因变量的 $n \times 1$ 向量, β 为待估参数,即回归系数的 $m \times 1$ 向量。这里的 n 为观测数, m 为待估计的回归系数(含截距项)的个数,其含义与第 2 章中的 p 略有不同。当 $X'X$ 矩阵不满秩或至少有一个特征根很小(即接近于零)时,则矩阵 X 为病态矩阵,此时用最小二乘法进行回归系数的估计就会产生较大的偏差,使 $\hat{\beta}$ 很不稳定,在具体取值上与真值有较大的偏差,甚至有时会出现系数的正负号与实际不符的情况。为此,Hoerl 和 Kennard 于 1970 年提出了岭估计(Ridge estimate)方法,即对多重线性回归方程的回归系数估计的方法如下:

$$\beta(k) = (X'X + kI_m)^{-1}X'Y \quad (21-2)$$

即在矩阵 $X'X$ 的主对角线元素上加上一个非负因子 k ,其中 I_m 为 m 阶单位矩阵, $k > 0$ 称为岭参数或偏参数。如果 k 取与试验数据 Y 无关的常数,则 $\beta(k)$ 为线性估计,否则 $\beta(k)$ 为非线性估计。取不同的 k ,得到不同的岭估计。所以上式实际上定义了一个很大的估计类。特别当 $k = 0$ 时, $\beta(k) = (X'X)^{-1}X'Y$ 就是 β 的最小二乘估计。当 $k \in [0, +\infty)$ 时,对于每个 i , $\beta(k)$ 的第 i 个分量 $\beta_i(k)$ 的值均为 k 的函数,在直角坐标系中,点 $(k, \beta_i(k))$ 所构成的变化轨迹,称为岭迹。随着 k 的增大, $\beta_i(k)$ 的绝对值均趋于不断变小(由于自变量之间可能存在

相关性，个别 $\beta_i(k)$ 可能有小范围的向上波动或改变正、负号)，若 $k \rightarrow +\infty$ ，则 $\beta(k) \rightarrow 0$ 。

与最小二乘估计相比，岭估计是把 $X'X$ 换成了 $X'X + kI$ 得到的。从直观上讲，当 X 为病态时， $X'X$ 的特征根至少有一个非常接近于 0，而 $X'X + kI$ 的特征根则变成 $\lambda_1 + k, \dots, \lambda_m + k$ ，它们中的某些接近于 0 的特征根就会得到改善，从而“打破”原来设计阵的复共线性，使得岭估计比最小二乘估计有较小的均方误差 ($MSE(\hat{\beta}(k)) < MSE(\beta)$)。

21.3 岭回归分析

21.3.1 进行多重线性回归分析并进行共线性诊断

设程序名为 SASTJFX21_1A.SAS，数据文件名为“DATA21_1.DAT”。程序如下：

```
data SASTJFX21_1;
infile "D:\SASTJFX\DATA21_1.DAT";
input id Y X1 -X4;
run;
ods html file = "D:\output21_1A.htm";

proc reg data = SASTJFX21_1;
model Y = X1 -X4 / COLLIN COLLINOINT;
run;
ods html close;
quit;
```

首先，建立数据集 SASTJFX21_1，读入数据，变量 Y、 $X_1 - X_4$ 分别代表血红蛋白含量及钙、镁、铁、铜的含量。然后，调用 REG 过程，“COLLIN”为对截距项未校正的共线性诊断，“COLLINOINT”为对截距项校正的共线性诊断。“ods html file = “D:\output21_1A.htm”；”利用 ODS 将 REG 过程产生的结果生成 HTML 格式文件，并指定文件存放于 D 盘根目录下，名称为 output21_1A.htm。

运行结果如下：

| Analysis of Variance | | | | | |
|----------------------|----|----------------|-------------|---------|--------|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 4 | 10667 | 2666.77653 | 24.25 | <.0001 |
| Error | 24 | 2639.34216 | 109.97259 | | |
| Corrected Total | 28 | 13306 | | | |

以上是对模型总体上是否具有统计学意义的方差分析结果，因 $F = 24.25$ 、 $P < 0.0001$ ，说明所求的回归模型总体上来说具有统计学意义。

| | | | |
|----------------|-----------|----------|--------|
| Root MSE | 10.48678 | R-Square | 0.8016 |
| Dependent Mean | 105.36207 | Adj R-Sq | 0.7686 |
| Coeff Var | 9.95309 | | |

| Parameter Estimates | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t |
| Intercept | 1 | 10.78224 | 15.13792 | 0.71 | 0.4832 |
| X1 | 1 | -2.87906 | 1.11230 | -2.59 | 0.0161 |
| X2 | 1 | 0.96252 | 1.25246 | 0.77 | 0.4497 |
| X3 | 1 | 1.60852 | 0.22664 | 7.10 | <.0001 |
| X4 | 1 | 82.38101 | 66.22720 | 1.24 | 0.2255 |

以上是对全模型参数进行最小二乘估计的结果,可以看出截距项 $t = 0.71$, $P = 0.4832 > 0.05$,说明截距项在 0.05 水平上无统计学意义,因此对自变量进行共线性诊断时应采用对截距项未做校正时的共线性诊断,其结果如下:

| Collinearity Diagnostics | | | | | | | |
|--------------------------|------------|-----------|-------------------------|------------|------------|------------|---------|
| Number | Eigenvalue | Condition | Proportion of Variation | | | | |
| | | Index | Intercept | X1 | X2 | X3 | X4 |
| 1 | 4.93078 | 1.00000 | 0.00067171 | 0.00054853 | 0.00048412 | 0.00062611 | 0.00110 |
| 2 | 0.03428 | 11.99292 | 0.05336 | 0.00728 | 0.00208 | 0.11458 | 0.48664 |
| 3 | 0.01752 | 16.77725 | 0.34637 | 0.15322 | 0.15430 | 0.16343 | 0.03290 |
| 4 | 0.00987 | 22.34885 | 0.35012 | 0.64321 | 0.14999 | 0.05482 | 0.44236 |
| 5 | 0.00755 | 25.55850 | 0.24948 | 0.19574 | 0.69315 | 0.66654 | 0.03700 |

以上是对截距项未做校正时的共线性诊断的结果。从最后一行的条件数可以看出 4 个自变量的最大条件数为 $25.558 > 10$,所以这 4 个自变量之间有较强的共线性。再从该条件数对应的方差分量上看,这 4 个自变量之间的共线性主要表现在 X_2 与 X_3 上(这两个变量的方差分量大于 0.5)。由此,可以认为全模型的设计矩阵是病态的,可以考虑用岭回归分析来进行参数估计。

| Collinearity Diagnostics(intercept adjusted) | | | | | | |
|--|------------|-----------|-------------------------|---------|---------|---------|
| Number | Eigenvalue | Condition | Proportion of Variation | | | |
| | | Index | X1 | X2 | X3 | X4 |
| 1 | 2.48779 | 1.00000 | 0.05636 | 0.04996 | 0.04283 | 0.05232 |
| 2 | 0.86649 | 1.69444 | 0.11835 | 0.02657 | 0.36324 | 0.12871 |
| 3 | 0.38251 | 2.55027 | 0.82483 | 0.06516 | 0.03152 | 0.47211 |
| 4 | 0.26320 | 3.07440 | 0.00046486 | 0.85831 | 0.56241 | 0.34685 |

以上是对截距项做校正后的共线性诊断的结果。本例因截距项在 0.05 水平上无统计学意义,因此对自变量进行共线性诊断时采用对截距项未做校正时的共线性诊断结果即可。

21.3.2 进行岭回归分析

对例 21-1 资料进行岭回归分析的 SAS 程序如下,设程序名为 SASTJFX21_1B.SAS。

```

symbol1 v=x c=blue;
symbol2 v=circle c=black;
symbol3 v=square c=red;
symbol4 v=triangle c=green;
legend1 position=(bottom
    right inside) mode=protect
across=3 cborder=
    black offset=(0,0)
label=(color=blue position=(top
center)' independent variables ');
cframe=white;
data SASTJFX21_1;
infile "D:\SASTJFX\DATA21_1.DAT";
input id Y X1 -X4;
run;

ods html file=
"D:\output21_1B.htm";
proc standard data=SASTJFX21_1
m=0 s=1 out=aa;
run;
proc reg data=aa outest=b
ridge=0 to 1.0 by .02;
model Y=X1 -X4;
plot /ridgeplot vref=0 lvref=1
nomodel legend=legend1 nostat;
run;
proc print data=b;
run;
ods html close;

```

程序的前10行为SAS中图形参数的设置,其中symbol1 - symbol4代表了4个自变量的符号及颜色,legend1后的参数为图例的设置,“position = (bottom right inside)”指定图例在图形的下、右、内侧,“mode = protect”表示图例出现在过程输出的区域,并且该图例区域没有其他的图形显示。“across = 3”指定图例条目中的列数为3,“cborder = black”指定图例区的边框的颜色为黑色。“label = (color = blue position = (top center) 'independent variables')”指定图例的标签,“cframe = white”指定填充图例区的颜色。

程序中的data步用于创建SAS数据集SASTJFX21_1。程序中的“standard”过程用于对数据进行标准化,并且将标准化的数据输出到新的数据集“aa”中去,“m = 0 s = 1”要求将数据集SASTJFX21_1中的所有变量标准化使得均值为0,标准差为1。“reg”过程中的“ridge = 0 to 1.0 by .02”表示岭参数 k 分别取0, 0.02, 0.04, ..., 1,并将参数估计的结果通过“outest = b”输出到数据集“b”中。“plot / ridgeplot”是绘制岭迹图,“vref = 0”指定垂直于垂直轴的参考直线出现的位置,“lvref = 1”指定用VREF = 选项要求的直线的类型为实线,“nomodel”不显示拟合的模型方程和标签,“nostat”在图形边沿不显示缺省时的统计量。通过“print”过程将这52个模型的参数打印出来。

运行结果如下:

在图21-1中,横坐标为 k 的取值,纵坐标为自变量的系数。可以看出当岭参数大于0.1时,自变量的回归系数趋于平稳(以水平直线为渐近线),但是 X_4 的系数虽然较稳定,但其绝对值始终较小,按照岭迹选变量的原则可将其从模型中去掉;由于 X_2 与 X_3 有较强的共线性,所以 X_2 与 X_3 只留一个即可,从图中看, X_2 的系数较小,应将其从模型中去掉。

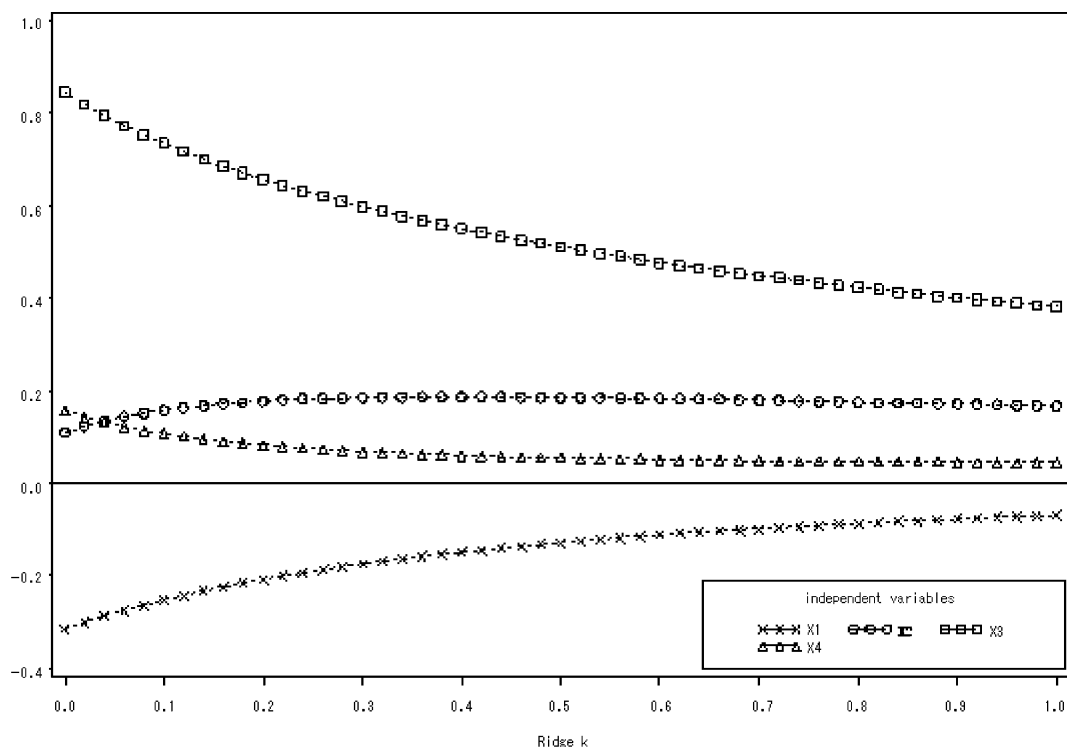


图 21-1 4 个自变量的系数随岭参数 k 的变化曲线图

| Obs | _MODEL_ | _DEP | _RIDGE_ | _RMSE_ | Intercept | x1 | x2 | x3 | x4 | y |
|-----|---------|------|---------|---------|-----------|----------|---------|---------|---------|-----|
| | | VAR_ | | | | | | | | |
| 1 | MODEL1 | y | . | 0.48105 | -8.74E-16 | -0.3149 | 0.11072 | 0.8462 | 0.1578 | -1 |
| 2 | MODEL1 | y | 0.00 | 0.48105 | -8.74E-16 | -0.3149 | 0.11072 | 0.8462 | 0.1578 | -1 |
| 3 | MODEL1 | y | 0.02 | 0.48168 | -8.38E-16 | -0.30016 | 0.12446 | 0.82013 | 0.14457 | -1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 50 | MODEL1 | y | 0.96 | 0.66679 | -3.06E-16 | -0.07134 | 0.17007 | 0.39021 | 0.04602 | -1 |
| 51 | MODEL1 | y | 0.98 | 0.6698 | -3.02E-16 | -0.06974 | 0.16923 | 0.38633 | 0.04581 | -1 |
| 52 | MODEL1 | y | 1.00 | 0.67276 | -2.98E-16 | -0.06818 | 0.16838 | 0.38253 | 0.0456 | -1 |

上表中是取不同的岭参数 k 时的回归系数的估计，一共有 52 个模型。将 X_2 与 X_4 从模型中去除后，重新绘制岭迹图，方法是将程序 SASTJFX21_1B.SAS 中的 model 语句改为“model Y = X1 X3/noint;”即可，重新运行程序，结果如下：

由图 21-2 可以看出，在 k 取 0.3 时，标准化回归系数趋于稳定，再由数据集“b”找出相应的模型参数估计。

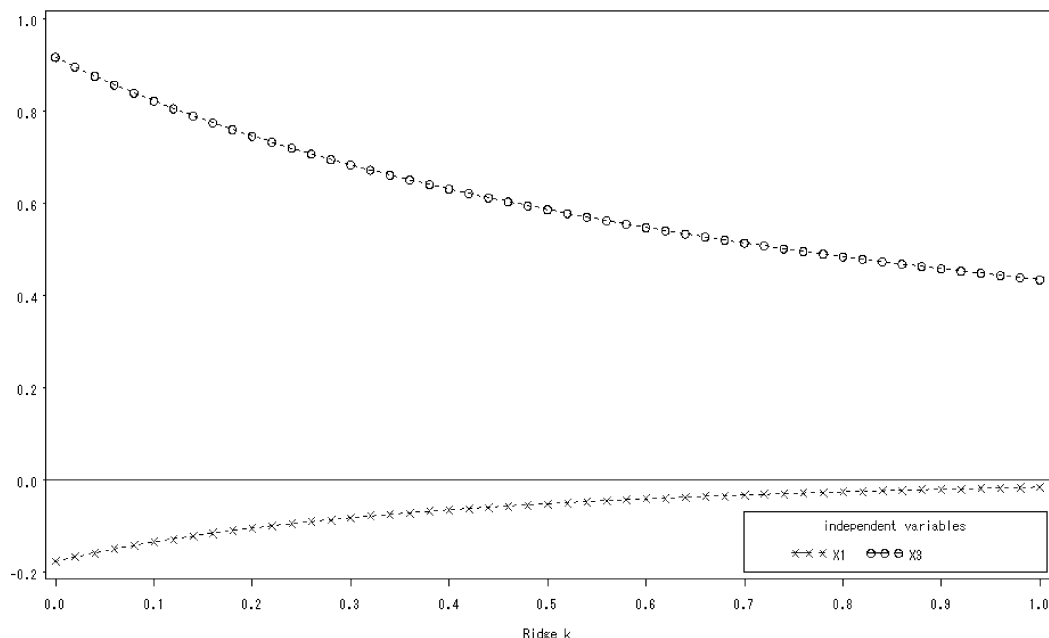


图 21-2 X_1 与 X_3 两个自变量的系数随岭参数 k 的变化曲线图

| | _MODEL_ | _TYPE_ | _DEPVAR_ | _RIDGE_ | _PCOMIT_ | _RMSE_ | X1 | X3 | Y |
|-----|---------|--------|----------|---------|----------|---------|----------|---------|-----|
| 1 | MODEL1 | PARMS | Y | . | . | 0.48403 | -0.17724 | 0.91658 | -1 |
| 2 | MODEL1 | RIDGE | Y | 0.00 | . | 0.49325 | -0.17724 | 0.91658 | -1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 16 | MODEL1 | RIDGE | Y | 0.28 | . | 0.54048 | -0.08651 | 0.69482 | -1 |
| 17 | MODEL1 | RIDGE | Y | 0.30 | . | 0.54528 | -0.08250 | 0.68320 | -1 |
| 18 | MODEL1 | RIDGE | Y | 0.32 | . | 0.55010 | -0.07870 | 0.67199 | -1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 51 | MODEL1 | RIDGE | Y | 0.98 | . | 0.68370 | -0.01712 | 0.43866 | -1 |
| 52 | MODEL1 | RIDGE | Y | 1.00 | . | 0.68679 | -0.01627 | 0.43415 | -1 |

岭参数为 0.3 时的标准化回归方程为：

$$\hat{Y} = -0.0825x_1 + 0.6832x_3$$

若将标准化后的变量还原成原变量,则需在岭回归程序中,将“data = aa”改为原始数据集“data = SASTJFX21_1”即可,在岭参数为 0.3 时的原变量的回归方程为:

回归方程为: $\hat{Y} = -0.75425X_1 + 1.29868X_3$ 。

从岭回归结果中可以得知:在 4 种微量元素中,铁与钙的含量对血红蛋白含量具有较重要的影响。从回归系数的符号可以得知铁的含量提高,有助于血红蛋白含量的提高,当钙的吸收量加大后,反而会使血红蛋白含量减少。

值得注意的是,如果该例采用逐步回归分析,方程为: $\hat{Y} = -1.24281X_1 + 1.81388X_3$ 。岭回归估计的方程与之相比系数的绝对值都变小了。从理论上讲,岭估计实际上将参数向量 $\hat{\beta}$ 的长度压缩到原来的 c 倍($0 < c < 1$),由此使残差平方和的上升尽可能小。在实际应用中,岭估计是一种应用很广泛的有偏估计,但是由于岭参数 k 的确定不是很容易的,所以不能简单地一概使用岭估计代替最小二乘估计。事实上,若设计矩阵不含共线性关系,最小二乘估计仍然是一个良好的估计,因此在应用时,只有当有充足的理由认为自变量之间有共线性时才适合应用岭估计,否则最好还是应用最小二乘估计建立多重线性回归方程。

21.4 与岭回归分析有关的 SAS 语句说明

进行岭回归分析时,其数据结构及录入方法与第 13 章中的多重线性回归分析相同,并且也是采用 SAS 软件中的 REG 过程进行分析。

在 SAS 软件中运用 REG 过程进行岭回归分析时,其对 k 的体现主要是通过 RIDGE 选项来实现的,并将估计值输出到 OUTEST = 指定的数据集中;例如:

```
proc reg data=aa outest=b ridge=0 to 1.0 by .02;
```

上述语句说明对名为“aa”的数据集进行岭回归分析,并将岭估计的参数值输入到数据集“b”当中。其中岭参数 k 在 0 至 1 的范围内,按步长为 0.02 取值,即 k 分别取 0, 0.02, 0.04, ..., 1。此时只是按不同的岭参数对数据进行岭估计,但是要绘制岭迹图还需在 REG 过程中的 PLOT 语句中加入 ridgeplot 选项。这时就可以根据岭迹图选择岭参数 k 了。语句格式如下:

```
plot /ridgeplot;
```

岭迹图可以显示随着岭参数 k 的变化,各自分量的变化趋势,选取趋于稳定态势时的岭参数对应的各自变量的回归系数,作为最后结果。还可将标准化变量还原为原变量来建立回归方程。

此外, SAS 软件中 GRAPH 模块中的设置参数也可以在此应用。如果使用 ODS GRAPHICS 功能,则不用 PLOT 语句就可直接输出岭迹图。

21.5 本章小结

本章主要介绍了岭回归分析的 SAS 实现方法。岭回归主要是当自变量之间存在共线性时应用,因为此时矩阵 $X'X$ 将会呈病态,若应用最小二乘法估计有可能会得出错误的结论。在进行岭回归分析时,岭参数 k 的估计非常重要。本章介绍了如何运用 SAS 软件中的 REG 过程来寻找合适的 k ,从而得出合理的模型。

(葛毅 周诗国)

第 22 章 Poisson 回归分析

Poisson 回归属于广义线性模型，专门适用于响应变量是计数资料的情形，它可以定量地分析多个影响因素与计数的响应变量之间的关系。本章将结合实例，阐述如何通过 SAS 软件实现 Poisson 回归的应用。

22.1 问题、数据及统计分析方法的选择

22.1.1 问题与数据

【例 22-1】 采用职业人群回顾性队列研究方法对所有从 1966 年 8 月 18 日到 1991 年 12 月 31 日在湖北某厂工作 5 年以上者的生存情况做了调查。符合进入队列的条件者有 9572 人，总共观察的人年数为 114488 人年，其中 159 人死亡，按年龄与是否暴露这两个因素分组的资料见表 22-1。试分析年龄与暴露因素对死亡率有无影响。

表 22-1 湖北某厂全死因死亡资料

| 年龄(岁) | 死亡数 | | 人年数 | | 死亡率(1/千) | |
|-------|-----|----|-------|-------|----------|--------|
| | 非暴露 | 暴露 | 非暴露 | 暴露 | 非暴露 | 暴露 |
| <40 | 39 | 30 | 59141 | 34995 | 0.659 | 0.857 |
| 40~49 | 14 | 33 | 6621 | 9241 | 2.114 | 3.571 |
| 50~59 | 3 | 25 | 650 | 3115 | 4.615 | 8.026 |
| 60~69 | 0 | 12 | 54 | 595 | 0.000 | 20.168 |
| ≥70 | 0 | 3 | 9 | 67 | 0.000 | 44.776 |

注：资料摘自字传华,余松林,向惠云. 中国卫生统计. 1996,13(1):6.

【例 22-2】 现有某医院在非器质性心脏病并且仅有胸闷症状的就诊者中随机收集 30 个患者在 24h 中的早搏数(y)。问：早搏是否与吸烟(x_1)、性别(x_2)和喝咖啡(x_3)有关？具体内容见表 22-2。

表 22-2 某医院非器质性心脏病伴有胸闷患者情况

| 编号 | x_1 | x_2 | x_3 | y | 编号 | x_1 | x_2 | x_3 | y | 编号 | x_1 | x_2 | x_3 | y |
|----|-------|-------|-------|----|----|-------|-------|-------|----|----|-------|-------|-------|----|
| 1 | 0 | 1 | 1 | 11 | 11 | 0 | 0 | 0 | 1 | 21 | 0 | 1 | 1 | 5 |
| 2 | 0 | 0 | 0 | 7 | 12 | 0 | 0 | 1 | 9 | 22 | 1 | 1 | 0 | 8 |
| 3 | 0 | 0 | 0 | 3 | 13 | 0 | 0 | 1 | 6 | 23 | 1 | 1 | 0 | 13 |
| 4 | 1 | 0 | 1 | 5 | 14 | 1 | 1 | 1 | 17 | 24 | 0 | 0 | 1 | 8 |
| 5 | 0 | 0 | 0 | 2 | 15 | 0 | 0 | 0 | 5 | 25 | 1 | 0 | 0 | 6 |
| 6 | 1 | 1 | 1 | 13 | 16 | 1 | 0 | 0 | 11 | 26 | 0 | 0 | 1 | 4 |
| 7 | 0 | 1 | 0 | 6 | 17 | 0 | 1 | 1 | 8 | 27 | 0 | 0 | 0 | 6 |
| 8 | 1 | 0 | 1 | 10 | 18 | 1 | 0 | 1 | 9 | 28 | 1 | 1 | 1 | 13 |
| 9 | 0 | 0 | 0 | 4 | 19 | 0 | 0 | 0 | 8 | 29 | 1 | 1 | 0 | 9 |
| 10 | 1 | 0 | 1 | 7 | 20 | 1 | 0 | 0 | 5 | 30 | 0 | 0 | 1 | 5 |

备注： $x_1=1$ 表示吸烟， $x_1=0$ 表示不吸烟； $x_2=1$ 表示喜欢喝咖啡， $x_2=0$ 表示不喜欢喝咖啡； $x_3=1$ 表示男性， $x_3=0$ 表示女性。

22.1.2 对数据结构的分析

实际研究中通常以数据库的形式呈现研究结果，每一个个体的观测用一行来表示，每一个（原因或结果）变量用一列来表示。这样的数据库呈现了研究资料的原始状态与结构，但不能明确地体现资料所对应的实验设计类型。为了清楚地体现资料所对应的实验设计类型，并方便介绍拟采用的统计学分析方法，同时突出资料特点与统计学分析方法的对应关系，常常采用某种从原始数据加工、提炼、整理而成的数据结构来呈现问题。因此，以下以分组和未分组两种情形的 Poisson 回归分析的数据结构的形式来呈现。在本章中，例 22-1 和例 22-2 即 Poisson 回归分析的数据结构的两种情形，对应的数据结构如表 22-3 和表 22-4 所示。

1. 未分组情形

表 22-3 从例 22-2 的实际问题归纳出的数据结构

| 编 号 | 自变量 1 (X_1) | 自变量 2 (X_2) | ... | 自变量 m (X_m) | 响应变量 (Y) |
|-----|-----------------|-----------------|-----|-----------------|--------------|
| 1 | x_{11} | x_{21} | ... | x_{m1} | y_1 |
| ... | ... | ... | ... | ... | ... |
| k | x_{1k} | x_{2k} | ... | x_{mk} | y_k |

注：响应变量在此服从 Poisson 分布。

2. 分组情形

表 22-4 从例 22-1 的实际问题归纳出的数据结构

| m 个自变量的水平组合 | 观测个体数 (n) | 响应变量 (Y) |
|-------------|-----------|--------------|
| 1 | n_1 | y_1 |
| ... | ... | ... |
| k | n_k | y_k |

注：m 个自变量的水平组合共有 k 种不同的情况，但每个自变量的水平数并不等于 k，每个自变量在 k 行上，至少有 2 个水平取值即可。

22.1.3 分析目的与统计分析方法的选择

定量地分析因变量与多个影响因素之间的关系时，通常可以考虑多重线性回归，但是在多重线性回归分析中，要求因变量服从正态分布，并且因变量与自变量之间的关系为线性关系。实际中的很多资料往往并不满足这些条件，例 22-1 和例 22-2 中的因变量都是计数的，很可能并不服从正态分布，而它们与自变量之间的关系用线性函数来描述也是不恰当的。这种情况下，就需要采用一个联接函数来表达因变量与自变量之间的关系，同时，考虑到因变量是计数的这一特点，假定因变量服从 Poisson 分布。此时，就可以采用 Poisson 回归对该资料进行分析。

例 22-1 中的资料是回顾性队列研究资料，该资料中每个人的观察时间不一定相同，故不能用基于二项分布的 logistic 回归分析来处理。该资料的特点是死亡率很低，如果假定人的死亡是相互独立的，则可认为死亡发生数服从 Poisson 分布。例 22-2 的资料中，由于患者早搏为个体计数资料，而且早搏较少，可认为近似服从 Poisson 分布。所以，上两例中的资料都可以考虑采用 Poisson 回归分析来拟合资料。

22.2 Poisson 回归分析应用场合及关键技术

22.2.1 Poisson 回归模型简介

Poisson 分布是由法国数学家(S. D. Poisson)作为二项分布的近似而引入的。对只有两个结果的 n 次独立重复随机试验, 当 n 很大, 且指定结果发生的概率 p 很小, 而 np 适中时, Poisson 分布是一个很好的近似。并且, Poisson 分布在描述稀有事件出现的概率时也特别有用, 例如, 某一城市的交通事故数, 某项保险的索赔次数, 放射性物质发射出的某种质点数, 在某种液体一个小体积中观察到的微生物数目等, 都可看成服从 Poisson 分布的随机变量。一般来说, Poisson 分布用来描述单位时间内或指定范围内特定事件出现次数的统计规律性。

设单位时间或空间内事件发生数为 k , k 的取值为 $0, 1, 2, \dots, \lambda$, 则概率分布

$$P\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots, \lambda \quad (22-1)$$

称为 Poisson 分布。

22.2.2 Poisson 回归定义

Poisson 回归是常用于单位时间、单位面积、单位空间内某事件发生数(即事件的发生服从 Poisson 分布)的影响因素分析的一种方法。

事实上, 在医学现象中有不少稀有事件的发生是符合 Poisson 分布的, 这些现象都可以选用 Poisson 回归模型来分析。

设 Y 是一个服从 Poisson 分布的随机变量, $X = (1, x_1, x_2, \dots, x_q)'$ 是一个协变向量, $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_q)'$ 是参数向量。如果 Y 的数学期望的对数可以表示为协变量的线性表达式, 即 $E(Y | X) = \exp(X'\beta)$, 则称 (Y, X) 服从泊松回归模型。对应的表达式为:

$$P(Y = k | X) = \frac{\lambda^k(X) e^{-\lambda(X)}}{k!}, \quad k = 0, 1, 2, 3, \dots \quad (22-2)$$

在广义线性模型中, 设因变量 Y 服从参数为 λ 的 Poisson 分布, 影响 λ 的因素为 x_1, x_2, \dots, x_m 。对服从 Poisson 分布的响应变量, 取联接函数为对数, 则:

$$\log(\lambda) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m \quad (22-3)$$

或写成:

$$\lambda = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m) \quad (22-4)$$

模型中假设各自变量对事件数的影响是指数相乘的, 故称为可乘效应的 Poisson 回归模型或 Poisson 乘法模型。回归系数 β_i 的解释是: 当其他自变量不变(不管取值是多少)时, 自变量 x_i 改变一个单位时, 平均事件数之对数值的改变量。

对于分组资料, 若各自变量组合之观察单位为 n_i , 则相应的发生数的估计值为:

$$\hat{y}_i = n_i \times \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m) \quad (22-5)$$

$\log(n_i)$ 称为偏移量。

22.2.3 Poisson 回归的参数估计

Poisson 回归的参数估计不能用最小二乘法估计,常用加权最小二乘法或极大似然法。具体计算步骤从略。需指出的是,对复杂资料来说,加权最小二乘估计与极大似然估计很难找到显式表达(即估计值可以用函数表达出来),均需用迭代法求解,常用的迭代法是 Newton-Raphson 迭代法。两种估计的结果也不尽相同,但往往是相近似的。极大似然估计有很多的优良性质。其中最重要的两条是:(1)对指数分布族中的分布,其极大似然估计是唯一的。(2)如果 $\hat{\theta}$ 是参数 θ 的极大似然估计, $g(\theta)$ 是 θ 的某函数,则 $g(\hat{\theta})$ 是 $g(\theta)$ 极大似然估计,这一性质称为不变性。因此,极大似然估计最为常用。

22.2.4 Poisson 回归估计系数的假设检验

(1)似然比检验:通过比较两个相嵌套模型(如模型 P 嵌套于模型 K 内)的对数似然函数统计量 G (又称 Deviance)来进行,其统计量为:

$$G = G_P - G_K = -2 \times (\text{模型 P 的对数似然函数} - \text{模型 K 的对数似然函数}) \quad (22-6)$$

其中,模型 P 中的变量是模型 K 中变量的一部分,另一部分就是我们要检验的变量。这里, G 服从自由度为 $K - P$ 的 χ^2 分布。

(2)回归系数的 Wald 检验:比较估计系数与 0 的差别是否有统计学意义,其检验统计量为:

$$z = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} \quad (22-7)$$

这里 z 为标准正态变量。参数的置信区间是基于 Wald 统计量导出的。

β 的 95% 置信区间为:

$$\hat{\beta} - 1.96 \times SE(\hat{\beta}) \sim \hat{\beta} + 1.96 \times SE(\hat{\beta}) \quad (22-8)$$

22.2.5 Poisson 拟合优度检验

1. Deviance 偏差统计量

设 $L(b_{\max}; y)$ 为全模型的似然函数, $L(b; y)$ 为选模型(即含部分自变量的模型)的似然函数,则定义 λ 统计量为:

$$\lambda = L(b_{\max}; y) / L(b; y) \quad (22-9)$$

计算 Deviance 偏差(D)统计量,采用如下公式:

$$D = 2 \ln \lambda = 2 [\ln L(b_{\max}; y) - \ln L(b; y)] \quad (22-10)$$

且 D 服从 χ^2 分布。很显然, D 越大,模型的估计值与观测值的偏差就越大,拟合效果就越差。

2. 广义 χ^2 统计量

$$\chi_G^2 = \sum \frac{(y - \hat{\mu})^2}{V(\hat{\mu})} \quad (22-11)$$

显然, χ_G^2 越大,估计值与观测值差别就越大,模型的拟合效果就越差。对正态分布来说, χ_G^2 就是离差平方和;对 Poisson 分布或二项分布来说, χ_G^2 就是一般的 Pearson χ^2 。

Deviance 偏差具有可加性,而 χ_G^2 不具有这种性质。但 χ_G^2 比 Deviance 偏差更易解释。

22.3 Poisson 回归分析的 SAS 实现

22.3.1 对例 22-1 资料进行分析

1. SAS 程序及说明

在 SAS 中, GENMOD 过程可实现 Poisson 回归模型的参数估计和拟合优度检验。

设对例 22-1 中资料拟合 Poisson 回归模型的程序名为 SASTJFX22_1.SAS, 程序如下:

| | |
|--|--|
| <pre>DATA SASTJFX22_1; INPUT AGE \$ EXPOSE N Y @@ ; AGE4 = (AGE = '40 -49 '); AGE5 = (AGE = '50 -59 '); AGE6 = (AGE = '60 -69 '); AGE7 = (AGE = ' >=70 '); CARDS; <40 0 59141 39 <40 1 34995 30 40 -49 0 6621 14 40 -49 1 9241 33 50 -59 0 650 3 50 -59 1 3115 25 60 -69 0 54 0 60 -69 1 595 12 >=70 0 9 0 >=70 1 67 3 ; RUN; ODS HTML; PROC GENMOD; MODEL Y/N = EXPOSE AGE4 - AGE7 / LINK = LOG DIST = nb noscale ; RUN;</pre> | <pre>PROC GENMOD; MODEL Y/N = EXPOSE AGE4 - AGE7 / LINK = LOG DIST = POI LRCI OBSTATS RESIDU- ALS TYPE1 TYPE3 scale = deviance; ODS OUTPUT ParameterEstimates = AA ObStats = CC; RUN; DATA BB (DROP = StdErr ChiSq Prob- ChiSq LowerLRCL UpperLRCL); SET AA; RR = EXP (Estimate); LCI = EXP (LowerLRCL); UCI = EXP (UpperLRCL); PROC PRINT; RUN; ODS HTML CLOSE;</pre> |
|--|--|

【程序说明】 AGE4 - AGE7 是将 AGE 产生哑变量, 在分析时以 AGE < 40 作为基线。“MODEL Y/N = EXPOSE AGE4 - AGE7”语句等号左边为发生数/总数, 右边为自变量。第一个 GENMOD 过程是用来检验数据是否存在过离散现象, 对是否存在过离散现象进行检验的方法有 Z 检验和 Lagrange 乘子检验, GENMOD 过程可以进行 Lagrange 乘子检验, 具体见本书第 23 章(负二项回归部分)。在该过程中, 通过拟合负二项回归模型, 并使用选项 NOSCALE(将冗余参数 k 的取值固定为 0, 此时相当于进行 Poisson 回归), 来输出 Lagrange 乘子检验的结果。

第二个 PROC GENMOD 过程中, MODEL 语句中选项 DIST = POISSON 指定资料中响应变量(误差项)的分布为 POISSON 分布, 选项 LINK = LOG 指定联接函数为对数函数。MODEL 语句后加 LRCI, 目的是为了输出参数估计值的置信区间, 通过置信区间的上下限值分别取自然对数的幂, 可得到相对危险度 RR 的置信区间。OBSTATS 是指输出预测值、变量的估计值及置信区间、粗残差、Pearson 残差、偏差残差、标化 Pearson 残差、似然比残差。TYPE1 选项要求给出第一型分析似然比统计量。TYPE3 选项要求给出第三型分析似然比统计量。BB 数据集的建立是为了由各变量估计值求出 RR 值及其 95% CI。

2. 主要分析结果及解释

以下是例 22-1 资料的输出结果。首先给出的是第一个 GENMOD 过程的部分输出结果。

第1部分：给出了 GENMOD 过程的模型信息。指出数据名称、误差分布、联接函数形式、响应变量、OFFSET 项和格子数以及两个分类变量的水平及取值。

| Model Information | |
|-----------------------------|-------------------|
| Data Set | WORK. SASTJFX22_1 |
| Distribution | Negative Binomial |
| Link Function | Log |
| Response Variable(Events) | Y |
| Response Variable(Trials) | N |
| Number of Observations Read | 10 |
| Number of Observations Used | 10 |
| Number of Events | 159 |
| Number of Trials | 114488 |

第2部分：输出 Lagrange 乘子统计量的结果。当在 MODEL 语句中定义了 NOSCALE 选项时，Lagrange 乘子统计量将被计算，检验 Poisson 模型中的过离散。这里检验统计量 $\chi^2 = 2766272.52$, $P < 0.001$ ，拒绝原假设，也就是拒绝了不存在过离散的假设，说明该资料存在过离散，此时若使用 Poisson 回归拟合该资料应对过离散进行校正。对过离散进行校正的方法有很多，可以对数据直接进行负二项回归分析，或采用相应的方法对过离散参数 ϕ 进行估计（过离散参数估计的方法有很多，具体参见相关著作）。进行 Poisson 回归分析时，参数的点估计值不会受过离散参数大小的影响，但是参数估计的标准误及置信区间会受到影响。在 SAS 中可以通过“SCALE = ”选项对过离散参数进行估计，如可选择“SCALE = DEVIANCE”（ $\phi = \text{离差} / \text{自由度}$ ）或“SCALE = PEARSON”（ $\phi = \chi^2 / \text{自由度}$ ）。值得注意的是，数据过离散现象产生的原因有很多，如数据间不独立、有重要的解释变量没有考虑进入模型、数据有异常值存在等等。因此，在选择模型拟合数据时，应仔细考虑数据的实际情况。

| Lagrange Multiplier Statistics | | |
|--------------------------------|------------|------------|
| Parameter | Chi-Square | Pr > ChiSq |
| Dispersion | 2766272.52 | < .0001 |

以下是第二个 GENMOD 过程输出的结果。

第3部分：给出模型拟合优度检验的结果。分别给出了偏差统计量、尺度化的偏差统计量、卡方值、尺度化卡方值和相应的自由度以及对数似然值，并显示计算收敛。本例 G^2 为 2.4927（ $\nu=4$ ），表明拟合良好。

| Criteria For Assessing Goodness Of Fit | | | |
|--|----|--------------|----------|
| Criterion | DF | Value | Value/DF |
| Deviance | 4 | 2.4927 | 0.6232 |
| Scaled Deviance | 4 | 4.0000 | 1.0000 |
| Pearson Chi-Square | 4 | 1.5422 | 0.3855 |
| Scaled Pearson X2 | 4 | 2.4747 | 0.6187 |
| Log Likelihood | | -185514.9747 | |

Algorithm converged.

第4部分：给出了参数估计值、参数标准误估计值、95% 置信区间、参数检验卡方值和 P 值。

Analysis Of Parameter Estimates

| Parameter | DF | Estimate | Standard Error | Likelihood Ratio 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|-----------|----|----------|----------------|--|---------|------------|------------|
| Intercept | 1 | -7.3903 | 0.1160 | -7.6254 | -7.1707 | 4062.31 | <.0001 |
| EXPOSE | 1 | 0.4086 | 0.1410 | 0.1352 | 0.6880 | 8.39 | 0.0038 |
| AGE4 | 1 | 1.3110 | 0.1521 | 1.0100 | 1.6067 | 74.28 | <.0001 |
| AGE5 | 1 | 2.1401 | 0.1860 | 1.7682 | 2.4970 | 132.42 | <.0001 |
| AGE6 | 1 | 3.0195 | 0.2557 | 2.4937 | 3.4940 | 139.42 | <.0001 |
| AGE7 | 1 | 3.7902 | 0.4698 | 2.7356 | 4.5828 | 65.08 | <.0001 |
| Scale | 0 | 0.7894 | 0.0000 | 0.7894 | 0.7894 | | |

Note: The scale parameter was estimated by the square root of DEVIANCE/DOF.

由此得出 Poisson 回归模型:

$$\begin{aligned} \text{Log}(\lambda) = & -7.3903 + 0.4086\text{Expose} + 1.3110\text{AGE4} \\ & + 2.1401\text{AGE5} + 3.0195\text{AGE6} + 3.7902\text{AGE7} \end{aligned}$$

该模型显示,暴露、年龄各因素均有统计学意义。死亡与暴露与否有关,暴露组的死亡风险是非暴露组的 1.5047 倍(RR 值);死亡与年龄有关,各年龄组对应的 RR 值显示,随着年龄的增大,死亡危险性逐渐增加。

基线的死亡率(即各自变量取值均为 0 时)为: $p = e^{-7.3903} = 0.62/\text{千}$ (95% CI: 0.0005 ~ 0.0008),由参数的估计值计算得到的 RR 值与 95% CI 如下:

| Obs | Parameter | DF | Estimate | RR | LCI | UCI |
|-----|-----------|----|----------|---------|---------|---------|
| 1 | Intercept | 1 | -7.3903 | 0.0006 | 0.0005 | 0.0008 |
| 2 | EXPOSE | 1 | 0.4086 | 1.5047 | 1.1447 | 1.9897 |
| 3 | AGE4 | 1 | 1.3110 | 3.7100 | 2.7457 | 4.9862 |
| 4 | AGE5 | 1 | 2.1401 | 8.5003 | 5.8603 | 12.1459 |
| 5 | AGE6 | 1 | 3.0195 | 20.4816 | 12.1064 | 32.9187 |
| 6 | AGE7 | 1 | 3.7902 | 44.2634 | 15.4189 | 97.7861 |
| 7 | Scale | 0 | 0.7894 | 2.2021 | 2.2021 | 2.2021 |

第 5 部分:给出了 TYPE1 和 TYPE3 分析的似然比统计量。第一型分析中,当引入暴露与否因素时,偏差统计量为 133.9757,再引入 AGE4 - AGE7 时分别为 115.2674、65.2691、19.3145、2.4927,模型的差别均有统计学意义,且拟合良好(第 3 部分已述)。在第三型的分析中,对各因素进行假设检验的结论与第一型的分析是一致的。

LR Statistics For Type 1 Analysis

| Source | Deviance | Num DF | Den DF | F Value | Pr > F | Chi-Square | Pr > ChiSq |
|-----------|----------|--------|--------|---------|--------|------------|------------|
| Intercept | 167.6069 | | | | | | |
| EXPOSE | 133.9757 | 1 | 4 | 53.97 | 0.0018 | 53.97 | <.0001 |
| AGE4 | 115.2674 | 1 | 4 | 30.02 | 0.0054 | 30.02 | <.0001 |
| AGE5 | 65.2691 | 1 | 4 | 80.23 | 0.0009 | 80.23 | <.0001 |
| AGE6 | 19.3145 | 1 | 4 | 73.74 | 0.0010 | 73.74 | <.0001 |
| AGE7 | 2.4927 | 1 | 4 | 26.99 | 0.0065 | 26.99 | <.0001 |

LR Statistics For Type 3 Analysis

| Source | Num DF | Chi-Square | Pr > ChiSq |
|--------|--------|------------|------------|
| EXPOSE | 1 | 8.62 | 0.0033 |
| AGE4 | 1 | 65.45 | <.0001 |
| AGE5 | 1 | 96.80 | <.0001 |

| | | | |
|------|---|-------|--------|
| AGE6 | 1 | 75.76 | <.0001 |
| AGE7 | 1 | 26.99 | <.0001 |

第 6 部分：输出预测值、粗残差、Pearson 残差、偏差残差、标化偏差残差、标化 Pearson 残差、似然比残差。

| | | | | | | | | | |
|-----|----|-------|--------|---------|---------|---------|----------|----------|---------|
| Obs | Y | N | Pred | Resraw | Reschi | Resdev | StResdev | StReschi | Reslik |
| 1 | 39 | 59141 | 0.0006 | 2.4992 | 0.4137 | 0.4092 | 1.1244 | 1.1367 | 1.1341 |
| 2 | 30 | 34995 | 0.0009 | -2.4992 | -0.4384 | -0.4444 | -1.1523 | -1.1367 | -1.1404 |
| 3 | 14 | 6621 | 0.0023 | -1.1606 | -0.2981 | -0.3023 | -0.5676 | -0.5596 | -0.5632 |
| 4 | 33 | 9241 | 0.0034 | 1.1606 | 0.2057 | 0.2048 | 0.5572 | 0.5596 | 0.5591 |
| 5 | 3 | 650 | 0.0052 | -0.4101 | -0.2221 | -0.2273 | -0.3231 | -0.3156 | -0.3216 |
| 6 | 25 | 3115 | 0.0079 | 0.4101 | 0.0827 | 0.0828 | 0.3160 | 0.3156 | 0.3157 |
| 7 | 0 | 54 | 0.0126 | -0.6826 | -0.8262 | -1.1721 | -1.5449 | -1.0889 | -1.5150 |
| 8 | 12 | 595 | 0.0190 | 0.6826 | 0.2029 | 0.2029 | 1.0889 | 1.0889 | 1.0889 |
| 9 | 0 | 9 | 0.0273 | -0.2459 | -0.4959 | -0.7061 | -0.9369 | -0.6579 | -0.9156 |
| 10 | 3 | 67 | 0.0411 | 0.2459 | 0.1482 | 0.1492 | 0.6627 | 0.6579 | 0.6583 |

整理成死亡率预测表，如表 22-5 所示。

表 22-5 湖北某厂全死因死亡率及其预测值

| 年龄(岁) | 死亡率(‰) | | 预测值(‰) | |
|-------|--------|--------|---------|---------|
| | 非暴露 | 暴露 | 非暴露 | 暴露 |
| <40 | 0.659 | 0.857 | 0.6172 | 0.9287 |
| 40~49 | 2.114 | 3.571 | 2.2898 | 3.4454 |
| 50~59 | 4.615 | 8.026 | 5.2462 | 7.8940 |
| 60~69 | 0.000 | 20.168 | 12.6409 | 19.0208 |
| ≥70 | 0.000 | 44.776 | 27.3186 | 41.1065 |

专业结论：暴露组相对于非暴露组死亡的相对危险度 $RR = 1.505$ ，说明暴露于特定环境增加了死亡的危险性。与小于 40 岁的年龄组相比，年龄增加 10 岁、20 岁、30 岁、30 岁以上时，死亡的相对危险度分别增加至 3.7100、8.5003、20.4816、44.2634(即为 AGE4、AGE5、AGE6、AGE7 的 RR 值)。

22.3.2 对例 22-2 资料进行分析

1. SAS 程序及说明

设对例 22-2 中资料拟合 Poisson 回归模型的程序名为 SASTJFX22_2.SAS，程序如下：

```
DATA SASTJFX22_2;
  INPUT x1 - x3 y @@ ;
CARDS;
0 1 1 11 0 0 0 7 0 0 0 3 1 0 1 5
0 0 0 2 1 1 1 13 0 1 0 6 1 0 1 10
0 0 0 4 1 0 1 7 0 0 0 1 0 0 1 9
0 0 1 6 1 1 1 17 0 0 0 5 1 0 0 11
0 1 1 8 1 0 1 9 0 0 0 8 1 0 0 5
0 1 1 5 1 1 0 8 1 1 0 13 0 0 1 8
1 0 0 6 0 0 1 4 0 0 0 6 1 1 1 13
1 1 0 9 0 0 1 5
;
RUN;
```

```
ODS HTML;
PROC GENMOD;
MODEL y = x1 - x3 / LINK = LOG DIST = NB
NOSCALE;RUN;
PROC GENMOD;
MODEL y = x1 - x3 / LINK = LOG DIST =
POISSON TYPE1 TYPE3
SCALE = DEVIANCE;
RUN;
ODS HTML CLOSE;
```

【程序说明】 该程序与例 22-1 中的程序大体相似，不同之处在于 MODEL 语句等号左侧只写出了因变量 Y，这是由于本例每个观测中早搏发生次数都是针对一例患者，可以理解为事件发生时基数都是相同的，另外还少了一些要求输出残差等统计量的选项。第一个 GENMOD 过程用来检验数据是否存在过离散现象，第二个 GENMOD 过程是对资料进行 POISSON 回归分析。

2. 主要分析结果及解释

以下是程序 SASTJFX22_2.SAS 的主要输出结果。首先给出第一个 GENMOD 过程的输出结果。

第 1 部分：给出了 GENMOD 过程的模型信息。指出数据名称、误差分布、联接函数形式、响应变量、OFFSET 项和格子数以及两个分类变量的水平及取值。

| Model Information | |
|-----------------------------|-------------------|
| Data Set | WORK. SASTJFX22_2 |
| Distribution | Negative Binomial |
| Link Function | Log |
| Dependent Variable | y |
| Number of Observations Read | 30 |
| Number of Observations Used | 30 |

第 2 部分：输出 Lagrange 乘子统计量的输出结果。这里检验统计量 $\chi^2 = 8.6315$ ， $P = 0.0033$ ，拒绝原假设，也就是拒绝了不存在过离散的假设，说明若使用 Poisson 回归拟合该资料应对过离散进行校正。

| Lagrange Multiplier Statistics | | |
|--------------------------------|------------|------------|
| Parameter | Chi-Square | Pr > ChiSq |
| Dispersion | 8.6315 | 0.0033 |

以下是第二个 GENMOD 过程的输出结果。

第 3 部分：给出模型拟合优度结果。分别给出了偏差统计量、尺度化的偏差统计量、卡方值、尺度化卡方值和相应的自由度以及对数似然值，并表示计算收敛。本例 G^2 为 23.5515 ($\nu = 26$)，表明拟合良好。

| Criteria For Assessing Goodness Of Fit | | | |
|--|----|----------|----------|
| Criterion | DF | Value | Value/DF |
| Deviance | 26 | 23.5155 | 0.9044 |
| Scaled Deviance | 26 | 26.0000 | 1.0000 |
| Pearson Chi-Square | 26 | 22.6916 | 0.8728 |
| Scaled Pearson X2 | 26 | 25.0890 | 0.9650 |
| Log Likelihood | | 265.7324 | |

第 4 部分：给出了参数、参数标准误估计值、95% 置信区间、参数检验卡方值和 P 值。可见 x1(吸烟否)、x2(喜欢喝咖啡否)、x3(性别)三个因素的系数与 0 的差异均有统计学意义 ($\beta_1 = 0.4162$, $P = 0.0015$; $\beta_2 = 0.4012$, $P = 0.0023$); $\beta_3 = 0.2546$, $P = 0.0493$)。

| Analysis Of Parameter Estimates | | | | | | | |
|---------------------------------|----|----------|----------------|----------|-------------------|------------|------------|
| Parameter | DF | Estimate | Standard Error | Wald 95% | Confidence Limits | Chi-Square | Pr > ChiSq |

| | | | | | | | |
|---|---|--------|--------|--------|--------|--------|--------|
| Intercept | 1 | 1.5066 | 0.1258 | 1.2600 | 1.7532 | 143.41 | <.0001 |
| x1 | 1 | 0.4162 | 0.1313 | 0.1588 | 0.6737 | 10.04 | 0.0015 |
| x2 | 1 | 0.4012 | 0.1314 | 0.1436 | 0.6588 | 9.32 | 0.0023 |
| x3 | 1 | 0.2546 | 0.1295 | 0.0008 | 0.5084 | 3.86 | 0.0493 |
| Scale | 0 | 0.9510 | 0.0000 | 0.9510 | 0.9510 | | |
| Note: The scale parameter was estimated by the square root of DEVIANCE/DOF. | | | | | | | |

第 5 部分: 给出了 TYPE1 和 TYPE3 分析的似然比统计量。在第一型的分析中, 当引入 x1 (吸烟否) 因素时, 偏差统计量为 36.4760, 再引入 x2 (喜欢和咖啡否) 时为 27.0489, 模型的差别都有统计学意义, 再引入 x3 (性别) 时为 23.5155, 模型的差别仍有统计学意义, 拟合良好 (第 3 部分已述)。在第三型的分析中, 对各因素进行假设检验的结论与第一型的分析是一致的。

| LR Statistics For Type 1 Analysis | | | | | | | |
|-----------------------------------|----------|------------|------------|---------|--------|------------|------------|
| Source | Deviance | Num DF | Den DF | F Value | Pr > F | Chi-Square | Pr > ChiSq |
| Intercept | 51.5146 | | | | | | |
| x1 | 36.4760 | 1 | 26 | 16.63 | 0.0004 | 16.63 | <.0001 |
| x2 | 27.0489 | 1 | 26 | 10.42 | 0.0034 | 10.42 | 0.0012 |
| x3 | 23.5155 | 1 | 26 | 3.91 | 0.0588 | 3.91 | 0.0481 |
| LR Statistics For Type 3 Analysis | | | | | | | |
| Source | Num DF | Chi-Square | Pr > ChiSq | | | | |
| x1 | 1 | 10.14 | 0.0015 | | | | |
| x2 | 1 | 9.20 | 0.0024 | | | | |
| x3 | 1 | 3.91 | 0.0481 | | | | |

专业结论: 以上结果说明吸烟者较不吸烟者、喜欢喝咖啡较不喜欢喝咖啡者、男性较女性易导致早搏数增加。

22.4 本章小结

本章介绍了如何进行 Poisson 回归分析。Poisson 回归用于描述结果变量服从 Poisson 分布的资料。Poisson 回归模型与 logistic 回归模型均属于广义线性模型, 在建模的过程中除了联接函数不同外, 主要的不同之处在于数据服从何种分布, 适合用何种方法建模。Poisson 回归一般用于单位时间、单位面积、单位空间内某事件发生数的影响因素的探讨, 事件的发生服从 Poisson 分布。当结果变量是二分类或多分类时, 应根据数据的分布情况看数据是满足 Poisson 分布还是可通过 logit 变换进行 logistic 回归分析。

Poisson 回归要求事件的发生是独立的。但医学研究中, 很多事件的发生是非独立的, 这种资料的特点是方差大于均数, 此时可选用负二项回归分析, 具体见本书有关章节。
(李长平)

第 23 章 负二项回归分析

本书有关章节中介绍了 Poisson 回归分析。负二项回归分析与 Poisson 回归分析一样，都属于广义线性模型，在模型的具体形式、参数估计、假设检验以及拟合优度检验方面有较多相似之处。本章将结合实例，讨论如何用 SAS 软件实现负二项回归分析。

23.1 问题、数据及统计分析方法的选择

23.1.1 问题与数据

【例 23-1】 本例中的数据来自于德国的卫生改革，改革的目的是减少病人到医生处的就诊次数。数据共包括 2227 例观测，分别属于改革之前的 1996 年和改革之后的 1998 年。响应变量为病人在三个月之内的就诊次数，自变量为是否改革(0 = 改革前, 1 = 改革后)、健康状况(0 = 良好, 1 = 不良)、年龄、受教育时间和家庭收入的对数值。由于数据量较大，仅列出部分结果，见表 23-1。

表 23-1 病人就诊次数及其影响因素数据

| 观测号 | 就诊次数 | 改革与否 | 健康状况 | 年 龄 | 受教育时间 | 家庭收入(取对数后) |
|------|------|------|------|-----|-------|------------|
| 1 | 1 | 0 | 0 | 45 | 10.5 | 7.6368 |
| 2 | 9 | 0 | 1 | 53 | 9.0 | 7.6992 |
| 3 | 40 | 0 | 1 | 48 | 10.5 | 7.0574 |
| 4 | 0 | 1 | 0 | 52 | 18.0 | 7.6886 |
| 5 | 1 | 0 | 0 | 40 | 10.5 | 7.5415 |
| 6 | 1 | 1 | 0 | 42 | 10.5 | 7.3319 |
| 7 | 0 | 0 | 1 | 57 | 10.5 | 7.4289 |
| 8 | 0 | 1 | 0 | 23 | 8.5 | 7.0978 |
| 9 | 3 | 0 | 0 | 55 | 10.0 | 7.8458 |
| ... | ... | ... | ... | ... | ... | ... |
| 2227 | 0 | 1 | 0 | 38 | 7.0 | 7.7232 |

【例 23-2】 关于某罕见病的调查研究，如表 23-2 所示，试以“发病人数”为因变量，采用 Poisson 回归和负二项回归分析该资料。

表 23-2 不同地区和不同年龄段某罕见病的发病情况

| 地 区 | 年龄(岁) | 调查人数 | 发病人数 | 地 区 | 年龄(岁) | 调查人数 | 发病人数 |
|-----|---------|--------|------|-----|---------|--------|------|
| A | 6 ~ 20 | 102250 | 0 | C | 6 ~ 20 | 87954 | 0 |
| | 21 ~ 35 | 89321 | 2 | | 21 ~ 35 | 90327 | 0 |
| | 36 ~ 50 | 76345 | 5 | | 36 ~ 50 | 78213 | 0 |
| B | 6 ~ 20 | 256789 | 26 | D | 6 ~ 20 | 123456 | 0 |
| | 21 ~ 35 | 223962 | 5 | | 21 ~ 35 | 112657 | 1 |
| | 36 ~ 50 | 199768 | 0 | | 36 ~ 50 | 108918 | 0 |

23.1.2 对数据结构的分析

例 23-1 中的数据结构与 Poisson 回归所对应的数据结构类似,因变量为就诊次数,属于计数的响应变量;自变量可以是定量变量,也可以是二值变量、多值有序变量或多值名义变量。表 23-1 中的改革与否和健康状况属于二值变量,年龄、受教育时间和取对数后的家庭收入是定量变量,这里是以数据库形式呈现的原始数据。

例 23-2 中的数据结构与 Poisson 回归所对应的数据结构类似,因变量为发病人数,属于计数的响应变量。显然,地区是一个多值名义变量,而年龄是一个多值有序变量,它们共有 $4 \times 3 = 12$ 种不同水平组合,以分组的形式呈现资料(即列联表形式的资料),不同自变量的每一种水平组合表现为一组,共有 12 组。

23.1.3 分析目的与统计分析方法的选择

对于例 23-1 中的资料也可以尝试采用 Poisson 回归分析,Poisson 回归分析是以计数资料为响应变量的标准回归模型。但是,Poisson 回归分析中要求均数和方差相等,实际数据往往并不符合这一假定,方差有时会大于均数,也就是所谓的过离散(overdispersion),这将导致模型参数估计值的标准误偏小,参数 Wald 检验的假阳性率增加。这种情况的出现可能是由于观测之间不独立导致的,而在医学研究中,很多事件的发生是非独立的。对于这类资料,可以采用负二项回归分析。过离散在理解负二项回归分析中居于中心地位,负二项回归的每一个应用几乎都与 Poisson 回归中发现过离散有关。

对于本例中的资料,首先使用 Poisson 回归进行拟合,然后根据拟合中遇到的问题,采用负二项回归来进行拟合。

对于例 23-2 中的资料也可以先尝试采用 Poisson 回归分析,然后,再采用负二项回归分析。此资料的分析留给读者去完成。

23.2 负二项回归分析应用场合及关键技术

负二项回归模型可由下式来表示:

$$g(\mu_i) = g(E(y_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im}, i = 1, 2, \cdots, n \quad (23-1)$$

该式与其他广义线性模型的形式(比如 Poisson 回归)相同,可以分为随机部分、系统部分与联接函数三个部分。随机部分是指响应变量以及它所对应的概率分布,这里指定响应变量服从负二项分布。系统部分是指等号右端的自变量部分,自变量的线性组合

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} \quad (23-2)$$

被称为线性预测值,该组合中的某些项可以是其中一些自变量的乘积项或者高次项,例如, $x_3 = x_1 x_2$ 或者 $x_4 = x_1^2$ 。联接函数 $g(\cdot)$ 将模型的随机部分与系统部分连接起来,它必须严格单调且充分光滑,即有足够阶数的导数。最简单的联接函数是恒等函数,即 $g(\mu) = \mu$,这正是经典的线性模型所采取的形式。联接函数也可以是非线性的,例如 logit 函数、probit 函数或者对数函数。对于每一种概率分布,都有一个均数的特殊函数,被称为自然参数。在广义线性模型中,使用自然参数作为联接函数称为典型联接,一般来说,典型联接在数学上处理起来比较方便,且有许多优良性质,最重要的优点是它使广义线性模型下统计推断的大样本理论更易处理。在实际中,典型联接的使用也最为普遍。在负二项回归中,典型联接为:

$$g(\mu) = \ln \frac{k\mu}{1 + k\mu} \quad (23-3)$$

负二项模型中参数的估计一般采用最大似然法, 其对数似然函数为:

$$\ln L = y \ln(k\mu) - (y + 1/k) \ln(1 + k\mu) + \ln \frac{\Gamma(y + 1/k)}{\Gamma(y + 1)\Gamma(1/k)} \quad (23-4)$$

参数的最大似然估计可以通过 Newton-Raphson 迭代算法获得, 最大似然法可以估计出冗余参数 k 。对参数的估计还可以使用加权最小二乘法, 它往往也要用迭代算法求解, 该法的缺点在于它不能估计冗余参数 k , k 必须作为已知常数被直接指定。

负二项回归中对所估计的参数进行检验的方法有似然比检验、Wald 检验和计分检验; 模型的拟合优度检验可以使用偏差统计量和广义 Pearson χ^2 统计量; 在对模型进行残差分析时, 常用的残差有 Pearson 残差、Anscombe 残差和偏差残差, 这些都与其他类型的广义线性模型相近, 负二项分布的偏差为:

$$\begin{aligned} D &= 2 \sum_{i=1}^n (\ln L(y; y) - \ln L(\mu; y)) \\ &= 2 \sum_{i=1}^n \left(y \ln(y/\mu) - (y + 1/k) \ln \frac{1 + ky}{1 + k\mu} \right) \end{aligned} \quad (23-5)$$

为了在更广泛的范围内拟合实际数据, 基本的负二项回归模型被扩展为一系列不同形式的模型。当方差 $V = \mu + k\mu$ 时, 称为 NB-1 模型; 当方差 $V = \mu + k\mu^p$ 时, 称为 NB-P 模型, 这里 p 被作为除 k 之外的另一个冗余参数。当参数 k 被设定为 1 时, 该模型就是几何模型, 与此相对应, 这时负二项分布退化为几何分布。在实际中, 一些计数资料并不包含计数为 0 的情形, 此时可以考虑采用零截尾(zero-truncated)负二项回归模型; 与之相反的情形是数据中包含了过多的零计数, 相应的改进后的模型为零堆积(zero-inflated)负二项回归模型。

零堆积负二项回归模型主要用于处理数据中包含过多零计数的情况, 其模型可以看成两个部分, 分别为二值模型部分和计数模型部分, 具体由分两个步骤来实现。对于第 i 个观测, 步骤一仅产生零计数, 它出现的概率为 φ_i ; 步骤二产生对应于负二项模型的计数, 其出现的概率为 $1 - \varphi_i$ 。可以由下式来表示:

$$y_i \sim \begin{cases} 0 & \text{概率为 } \varphi_i \\ f(y_i) & \text{概率为 } 1 - \varphi_i \end{cases}$$

其中 $f(y_i)$ 为负二项分布的概率函数, 如果拟合零堆积 Poisson 回归模型, 此处只要使用 Poisson 分布的概率函数即可。因此, 对于第 i 个观测, 其取值的概率为:

$$\begin{aligned} P(y_i = 0 \mid \mathbf{x}_i) &= \varphi_i + (1 - \varphi_i)f(0) \\ P(y_i \mid \mathbf{x}_i) &= (1 - \varphi_i)f(y_i), y_i > 0 \end{aligned} \quad (23-6)$$

其中 \mathbf{x}_i 为自变量。由上式可以看到, 一个观测取值为 0 的概率分为两部分: 其一为步骤一出现的概率, 其二为步骤二出现的概率与负二项模型中产生 0 的概率之积。

接着就可以对概率 φ_i 建模, 它可以被表示为:

$$\varphi_i = F_i = F(\gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_m z_{im}), i = 1, 2, \cdots, n \quad (23-7)$$

其中 $\gamma_0, \gamma_1, \cdots, \gamma_m$ 是待估计的零堆积系数; z_1, z_2, \cdots, z_m 为零堆积自变量; F_i 为零堆积联接函数, 这个函数可以是 logistic 函数, 也可以是 probit 函数, SAS 中使用这两种联接函数。

Vuong 检验统计量可以用于比较负二项回归模型与零堆积负二项模型的适宜性, 当然, 它也同样可以用于 Poisson 回归模型与零堆积 Poisson 模型。检验统计量 V 渐近服从正态分布,

如果 $V > 1.96$, 说明应该使用零堆积模型; 如果 $V < -1.96$, 应使用一般的负二项模型; 如果它的取值在 $-1.96 \sim 1.96$ 之间, 则这两种模型都不合适。

23.3 负二项回归分析的 SAS 实现

23.3.1 SAS 程序及说明

设对例 23-1 中资料拟合 Poisson 回归模型的程序名为 SASTJFX23_1A.SAS, 程序如下:

| | |
|---|--|
| <pre>Data SASTJFX23_1A; infile 'd:\SASTJFX\mdvisit.dat'; input id numvisit reform badh age edu loginc; run; ODS HTML;</pre> | <pre>proc genmod data = SASTJFX23_1A; model numvisit = reform badh age edu loginc/dist = poisson; run; ODS HTML CLOSE;</pre> |
|---|--|

【程序说明】 首先建立数据集 prg23_1, 读入数据, 变量 id、numvisit、reform、badh、age、edu、loginc 分别表示观测号、就诊次数、是否改革、健康状况、年龄、受教育时间和家庭收入的对数值。然后调用 GENMOD 过程, MODEL 语句中使用选项 DIST = POISSON 指定拟合 Poisson 回归模型。

在拟合 Poisson 回归模型的基础上, 继续对该资料拟合负二项回归模型, 程序名为 SASTJFX23_1B.SAS, 程序如下:

| | |
|--|---|
| <pre>data SASTJFX23_1B; infile 'd:\SASTJFX\mdvisit.dat'; input id numvisit reform badh age edu loginc; run; ODS HTML; proc genmod data = SASTJFX23_1B; model numvisit = reform badh age edu loginc/dist = nb link = log noscale;</pre> | <pre>run; proc genmod data = SASTJFX23_1B; model numvisit = reform badh age edu loginc/dist = nb link = log type1 type3; run; ODS HTML CLOSE;</pre> |
|--|---|

【程序说明】 程序分为三部分。第一部分为数据步, 同样是建立数据集 prg23_1。程序的第二部分, 也就是第一个 GENMOD 过程步, 拟合负二项回归模型, 与 Poisson 回归不同的是, 这里 DIST 选项的取值为 NB。需要特别指出的是此处使用了选项 NOSCALE, 是将冗余参数 k 的取值固定为 0, 此时相当于进行 Poisson 回归, 同时可以输出 Lagrange 乘子检验的结果。程序的第三部分同样拟合负二项回归, 选项 TYPE1 要求进行第一型或顺序的分析, 选项 TYPE3 要求为 MODEL 语句中规定的每个效应计算有关第三型对比的统计量。在 MODEL 语句中, 其他一些选项还可输出均数的预测值、线性预测值的估计值以及不同类型的残差, 具体可参见第 22 章。

23.3.2 主要分析结果及解释

首先给出程序 SASTJFX23_1A.SAS 的输出结果, 也就是对例 23-1 中资料拟合 Poisson 回归的结果, 由于第 22 章已有关于 Poisson 回归结果的详细说明, 故一些次要的输出结果都略去, 仅给出如下主要输出结果:

The GENMOD Procedure
Criteria For Assessing Goodness Of Fit

| Criterion | DF | Value | Value/DF |
|--------------------|------|-----------|----------|
| Deviance | 2221 | 7422.1244 | 3.3418 |
| Scaled Deviance | 2221 | 7422.1244 | 3.3418 |
| Pearson Chi-Square | 2221 | 9681.6920 | 4.3592 |
| Scaled Pearson X2 | 2221 | 9681.6920 | 4.3592 |
| Log Likelihood | | 432.6923 | |

Analysis Of Parameter Estimates

| Parameter | DF | Estimate | Standard Error | Wald 95% | Confidence Limits | Chi-Square | Pr > ChiSq |
|-----------|----|----------|----------------|----------|-------------------|------------|------------|
| Intercept | 1 | -0.4215 | 0.2690 | -0.9487 | 0.1057 | 2.46 | 0.1171 |
| reform | 1 | -0.1399 | 0.0265 | -0.1919 | -0.0878 | 27.76 | <.0001 |
| badh | 1 | 1.1326 | 0.0303 | 1.0732 | 1.1920 | 1397.43 | <.0001 |
| age | 1 | 0.0049 | 0.0013 | 0.0024 | 0.0073 | 15.21 | <.0001 |
| edu | 1 | -0.0118 | 0.0060 | -0.0235 | -0.0001 | 3.93 | 0.0474 |
| loginc | 1 | 0.1520 | 0.0360 | 0.0815 | 0.2226 | 17.85 | <.0001 |
| Scale | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 | | |

以上是模型拟合优度和参数估计的结果，表面上看，该模型似乎拟合较好，对自变量回归系数进行检验都是有统计学意义的。然而，由于 Pearson 离散统计量的值，也就是 Pearson χ^2 值与其自由度之比为 4.3592，说明数据可能存在过离散。Pearson 离散统计量被定义为 Pearson χ^2 值与其自由度之比，它可以说明 Poisson 回归和负二项回归模型中的过离散，反映数据潜在的变异。当该统计量大于 1.0，说明该模型可能是过离散的。该统计量不太大时，可以不用太多关注；然而，对于中等程度大小的模型而言，如果该统计量大于 1.25，就需要进行校正；对于大样本资料，离散统计量超过 1.05 就意味着可能存在过离散。该统计量小于 1 的情形被称为过低离散 (underdispersed)。偏差 (Deviance) 与其自由度的比值同样也可以衡量过离散。偏差和 Pearson χ^2 值与对应自由度的比值可以作为 Poisson 回归中离散参数的估计值。

过离散是否显著依赖于三点：Pearson 离散统计量的取值、模型中观测的数目和数据结构。负二项回归模型的一些改进形式，如零截尾负二项回归和零堆积负二项回归模型等可以用来解决负二项模型中的过离散问题，而负二项回归本身，就是解决 Poisson 回归中数据过离散的方法之一。事实上，正确认识这几个模型之间的关系是非常必要的。

对是否存在过离散进行检验的方法有 Z 检验和 Lagrange 乘子检验，GENMOD 过程没有提供 Z 检验的结果，却可以进行 Lagrange 乘子检验，在后面的输出结果中将进行详细介绍。

接着给出程序 SASTJFX23_1B.SAS 的输出结果，也就是对例 23-1 中资料拟合负二项回归的结果。

Lagrange Multiplier Statistics

| Parameter | Chi-Square | Pr > ChiSq |
|------------|------------|------------|
| Dispersion | 551.7077 | <.0001 |

首先输出的是第一个 GENMOD 过程步所产生的结果，由于其他部分与拟合 Poisson 回归时结果一致，故略去，只给出 Lagrange 乘子统计量的输出结果。当在 MODEL 语句中定义了 NOINT 或 NOSCALE 选项时，Lagrange 乘子统计量将被计算，该统计量用来评价对于截距或尺度参数所施加限定的有效性，渐近服从自由度为 1 的 χ^2 分布。在负二项回归中，设定

NOSCALE 选项时, 该统计量可以检验 Poisson 模型中的过离散。这里检验统计量 $\chi^2 = 551.7077$, $P < 0.001$, 拒绝原假设, 也就是拒绝了不存在过离散的假设, 说明使用 Poisson 回归拟合该资料是不合适的。

| The GENMOD Procedure | |
|-----------------------------|-------------------|
| Model Information | |
| Data Set | WORK.PR23_1 |
| Distribution | Negative Binomial |
| Link Function | Log |
| Dependent Variable | numvisit |
| Number of Observations Read | 2227 |
| Number of Observations Used | 2227 |

以上是第二个 GENMOD 过程步输出的第一部分结果, 是关于模型的一些基本信息, 其中数据集名称为 PRG23_1, 模型中使用负二项分布, 联接函数是对数函数, 响应变量为 numvisit, 观测数为 2227 个。

| Criteria For Assessing Goodness Of Fit | | | |
|--|------|-----------|----------|
| Criterion | DF | Value | Value/DF |
| Deviance | 2221 | 2412.3440 | 1.0862 |
| Scaled Deviance | 2221 | 2412.3440 | 1.0862 |
| Pearson Chi-Square | 2221 | 2693.5477 | 1.2128 |
| Scaled Pearson X2 | 2221 | 2693.5477 | 1.2128 |
| Log Likelihood | | 1813.1299 | |

Algorithm converged.

第二部分结果是评价模型拟合优度的统计量, 依次为偏差、尺度化的偏差、Pearson χ^2 值、尺度化 Pearson χ^2 值和对数似然函数值。可以看到, 与 Poisson 回归模型相比, 偏差从 7422.1244 减少到 2412.3440, Pearson χ^2 从 9681.6920 减少到 2693.5477。这两个值越大, 说明模型拟合得越差。根据 Poisson 回归模型与负二项回归模型在这两个统计量上的取值大小关系, 可以说明: 就本资料而言, 负二项回归模型的拟合效果要优于 Poisson 回归模型。此外, Pearson χ^2 离散统计量的取值也从 4.3592 下降到 1.2128。

| Analysis Of Parameter Estimates | | | | | | | |
|---------------------------------|----|----------|----------------|----------|-------------------|------------|------------|
| Parameter | DF | Estimate | Standard Error | Wald 95% | Confidence Limits | Chi-Square | Pr > ChiSq |
| Intercept | 1 | -0.4075 | 0.5336 | -1.4533 | 0.6384 | 0.58 | 0.4451 |
| reform | 1 | -0.1374 | 0.0511 | -0.2376 | -0.0372 | 7.22 | 0.0072 |
| badh | 1 | 1.1312 | 0.0748 | 0.9847 | 1.2778 | 228.84 | <.0001 |
| age | 1 | 0.0056 | 0.0024 | 0.0009 | 0.0103 | 5.39 | 0.0203 |
| edu | 1 | -0.0053 | 0.0115 | -0.0278 | 0.0172 | 0.21 | 0.6430 |
| loginc | 1 | 0.1371 | 0.0712 | -0.0024 | 0.2767 | 3.71 | 0.0541 |
| Dispersion | 1 | 1.0021 | 0.0476 | 0.9088 | 1.0955 | | |

The negative binomial dispersion parameter was estimated by maximum likelihood.

第三部分是参数估计的结果, 包括各个参数所对应的自由度、估计值、标准误、95% 可信区间、Wald χ^2 值和 P 值。其中截距项的估计值为 -0.4075, 对它进行假设检验的 P 值为 0.4451, 说明截距项与 0 之间的差别没有统计学意义。是否改革、健康状况和年龄这三个自变量的系数估计值分别为 -0.1374、1.1312 和 0.0056, 经检验 P 值都小于 0.05, 说明有统计学

意义。受教育时间和取对数后的家庭收入对应的系数估计值分别为 -0.0053 和 0.1371 ，都没有统计学意义。除了截距和各自变量的系数以外，此部分结果中还包含对于负二项过离散参数 k 的估计，这里估计值为 1.0021 ，标准误为 0.0476 ，其 95% 可信区间为 $(0.9088, 1.0955)$ 。当参数 k 被设定为 1 时，负二项模型就是几何模型，从过离散参数估计的结果来看，它非常接近于 1，并且其 95% 可信区间也包含 1，所以该模型实际上也就是一个几何模型。

由以上结果可以得到负二项回归模型的方程为：

$$\text{Log}(\mu) = -0.4075 - 0.1374\text{reform} + 1.1312\text{badh} + 0.0056\text{age} - 0.0053\text{edu} + 0.1371\text{loginc}$$

LR Statistics For Type 1 Analysis

| Source | 2 * LogLikelihood | DF | Chi-Square | Pr > ChiSq |
|-----------|-------------------|----|------------|------------|
| Intercept | 3325.8011 | | | |
| reform | 3337.2855 | 1 | 11.48 | 0.0007 |
| badh | 3615.7450 | 1 | 278.46 | <.0001 |
| age | 3622.5601 | 1 | 6.82 | 0.0090 |
| edu | 3622.5602 | 1 | 0.00 | 0.9936 |
| loginc | 3626.2597 | 1 | 3.70 | 0.0544 |

LR Statistics For Type 3 Analysis

| Source | DF | Chi-Square | Pr > ChiSq |
|--------|----|------------|------------|
| reform | 1 | 7.21 | 0.0073 |
| badh | 1 | 247.93 | <.0001 |
| age | 1 | 5.40 | 0.0202 |
| edu | 1 | 0.21 | 0.6433 |
| loginc | 1 | 3.70 | 0.0544 |

第四部分结果是由 MODEL 语句中的选项 TYPE1 和 TYPE3 产生的结果，分别是第一型和第三型分析似然比统计量。在第一型分析的结果中，包括 2 倍的对数似然函数值、自由度、 χ^2 值和 P 值。对于是否改革、健康状况和年龄检验的 P 值分别为 0.0007、小于 0.0001 和 0.0090，都是有统计学意义的，说明这三个效应的作用具有显著性。而受教育时间和取对数后的家庭收入对应的 P 值分别是 0.9936 和 0.0544，没有统计学意义。在第三型的分析中，对各因素进行假设检验的结论与第一型的分析是一致的。

专业结论：综合以上分析可以看出，改革与否、健康状况和年龄这三个因素对病人的就诊次数存在影响，而就现有的数据来看，还无法得出受教育时间和取对数后的家庭收入对就诊次数有影响的结论。结合参数估计的结果来看，改革与否的系数为 -0.1374 ，说明改革之后，病人的就诊次数下降了，由 $\exp(-0.1374) = 0.8716$ 可知，在不考虑其他因素的基础上，改革之后与改革前相比，就诊次数减少了大约 12.84%；而健康状况与年龄的参数估计值均大于 0，说明随着年龄的增加，就诊次数有增加的趋势，而不良的健康状况也会导致就诊次数的增加。本研究中最主要的目的就是考察改革是否减少了病人到医生处的就诊次数，综上所述，这项改革实现了相应的目标。

23.4 GENMOD 过程及 COUNTREG 过程简介

负二项回归可以由 SAS/STAT 模块的 GENMOD 过程完成，基本的程序语句与 Poisson 回归等其他广义线性模型大体上相似，不同之处是在 MODEL 语句中指定分布类型时，需要指定为

负二项分布，具体语句格式为：

```
MODEL response = <effects> / DIST |D| ERROR |ERR = NEGBIN |NB;
```

其中 DIST、D、ERROR 和 ERR 是等价的，NEGBIN 与 NB 是等价的，用“|”号隔开表示任选其一。在该过程中，对应于负二项分布的默认联接函数是对数函数，在 MODEL 语句后不使用 LINK 选项指定联接函数类型时使用该函数。

另外需要特别说明的是 GENMOD 过程中 VARIANCE 和 DEVIANCE、FWDLINK 与 INV-LINK 这四个语句。除了 GENMOD 过程内置的概率分布以外，用户可以根据需要使用 VARIANCE 和 DEVIANCE 规定一个概率分布，VARIANCE 语句定义方差函数，DEVIANCE 语句定义偏差函数；使用 FWDLINK 语句定义联接函数，INVLINK 语句定义联接函数的反函数，其类型并不仅仅局限于 SAS 本身所提供的一些函数形式。它们的语句格式为：

```
VARIANCE variable = expression;
DEVIANCE variable = expression;
FWDLINK variable = expression;
INVLINK variable = expression;
```

这四个语句的格式基本相同，语句中 variable 标识过程中对应的函数，expression 定义函数的表达式，它可以是 DATA 步的 SAS 语言所支持的任何数学表达式。在表达式中，可以使用自动变量 _MEAN_、_RESPONSE_ 和 _XBETA_，它们分别代表均数、响应变量和线性预测值。

上述四个函数都可以使用程序设计语句定义，在程序设计语句中指定一个变量代表该函数，最后将该变量列在 expression 的位置。

另外，当用 VARIANCE 语句定义方差函数，DEVIANCE 语句也必须同时出现；如果用 FWDLINK 语句定义联接函数，也必须同时使用 INVLINK 语句指定其反函数。

关于 GENMOD 过程语句的具体说明，可以参阅 SAS 说明书及相关文献，此处不再赘述。

在 SAS 9.2 版本中，ETS 模块的 COUNTREG 过程也可以完成 Poisson 回归和负二项回归，同时也可以实现零堆积 Poisson 回归和零堆积负二项回归，其语句格式为：

```
PROC COUNTREG options;
  BOUNDS bound1 [,bound2...];
  BY variables;
  INIT initvalue1 [,initvalue2...];
  MODEL dependent variable = regressors/options;
  OUTPUT options;
  RESTRICT restriction1 [,restriction2...];
  ZEROMODEL dependent variable ~ zero-inflated regressors/options;
```

PROC COUNTREG 语句说明执行 COUNTREG 过程，它后面的选项包括数据集选项、输出数据集选项、打印选项、估计控制选项和控制最优化过程的选项，其中控制最优化过程的选项格式为 METHOD = value，可以选择的迭代方法有准 Newton 法、Newton-Raphson 法和信赖域法 (trust region method)，分别用 QN、NRA 和 TR 表示。

BOUNDS 语句用来限定参数估计值的范围，它可以通过不等式来表示参数的取值范围，例如，下列语句

```
BOUNDS z < 0, 0 < x1 - x10 < 1;
```

表示将参数 z 的估计值限定为非负，同时将参数 x1 至 x10 的估计值限定在 0 ~ 1 范围内。与可以完成同样功能 RESTRICT 语句相比，BOUNDS 语句更为简洁。

BY 语句定义一个分组变量,按该变量分组后在各组内单独进行分析。

INIT 语句为参数初始值赋值语句,在使用迭代算法估计参数值时,需要根据初始值来进行后面的迭代计算。

MODEL 语句中等号左侧为响应变量,右侧为自变量。响应变量的取值应该是非负整数,当数据中某个观测的响应变量取负值时,该观测将被舍弃。该语句后的选项包括 DIST 选项、NOINT 选项、OFFSET 选项和打印选项。DIST 选项定义模型类型,它的语句格式为 DIST = value 或 D = value,模型种类包括 POISSON|P、NEGBIN(P = 1)、NEGBIN(P = 2)|NEGBIN、ZI-POISSON|ZIP 和 ZINEGBIN|ZINB,分别代表 Poisson 回归模型、含线性方差函数的负二项回归模型、含二次方差函数的负二项回归模型、零堆积 Poisson 回归模型和零堆积负二项回归模型。NOINT 选项规定去掉截距项。OFFSET 选项指定输入数据集中的—个变量作为偏移变量,该变量不能是响应变量、自变量或零堆积偏移变量。

OUTPUT 语句的用法与其他过程相近,可参考其他章节的叙述。

RESTRICT 语句的用法与 BOUNDS 语句有相似之处,它对参数估计值施加线性约束,每个限制条件被写作用等号或不等号连接的表达式,表达式必须是变量的线性函数。语句

```
RESTRICT_alpha = 1;
```

表示过离散参数 k 的取值为 1,此时负二项分布的方差为 $\mu + \mu^2$ 。

如果在 MODEL 语句中定义了零堆积 Poisson 回归模型或零堆积负二项回归模型,就必须使用 ZEROMODEL 语句,该语句中的响应变量必须与 MODEL 语句里的响应变量一致。该语句的选项有 LINK 选项和 OFFSET 选项,选项 LINK 定义用于计算取值为 0 的概率的分布函数,可选择的分布有标准正态分布和 logistic 分布,分别用 NORMAL 和 LOGISTIC 表示;OFFSET 选项指定零堆积偏移变量,该变量不能是响应变量、自变量或 MODEL 语句中的偏移变量。

23.5 本章小结

本章介绍了负二项回归分析的用法及其 SAS 实现方法。属于广义线性模型的负二项回归模型是基于负二项分布的,理解这个模型的关键在于掌握它与 Poisson 回归模型的联系。负二项回归可以解决 Poisson 回归中存在的过离散问题,将方差表示为均数的二次函数的形式。在 SAS 软件中,GENMOD 过程与 COUNTREG 过程均可用于实现负二项回归分析。

(柳伟伟 周诗国 陶丽新)

第 24 章 Probit 回归分析

Probit(概率单位)回归分析与 logistic 回归分析一样,都是在研究一个响应变量与多个自变量(可以是连续型变量或离散型变量)之间的关系中,响应变量为二值变量时可能用到的方法。它们之间的区别是:Probit 回归分析所适用的资料是“以分组形式”呈现的,即全部自变量的任何一种组合(属于一个特定的小组)下反应变量必须有一个研究者关心的某事件的“阳性率”数据;而 logistic 回归分析可用于以“个体”或“组”为观察对象的资料中,但其结局必须是定性结果变量的某一种水平出现或不出现的标志,例如,若是二值结果变量,出现阳性或阴性;若是多值结果变量,出现多值中的哪一种。简而言之,logistic 回归分析强调的是特定条件下所关心事件究竟出现哪一种具体情况;而 Probit 回归分析强调的是特定条件下所关心事件究竟出现的概率有多大。本章将结合实例,讨论如何用 SAS 软件实现 Probit 回归分析。

24.1 问题、数据及统计分析方法的选择

24.1.1 问题与数据

【例 24-1】 某研究者研究打鼾与罹患心脏病的关系。采用问卷调查,透过 2484 位受访者的配偶指出受访者夜晚打鼾的情况,分别以不曾打鼾、偶尔打鼾、几乎每晚打鼾及每晚打鼾来做区别,并对各种情形给以一个评分,得到如表 24-1 所示的结果。请问:打鼾的程度与罹患心脏病的风险是否有关?若有关,可否根据表 24-1 资料建立罹患心脏病的风险与打鼾程度的定量关系?如何建立?

表 24-1 打鼾与罹患心脏病情况的调查结果

| 打鼾情形 | 对打鼾情形的评分 | 人 数 | | |
|--------|----------|------------|----|------|
| | | 是否罹患心脏病: 否 | 是 | 合计 |
| 不曾打鼾 | 0 | 1355 | 24 | 1379 |
| 偶尔打鼾 | 2 | 603 | 35 | 638 |
| 几乎每晚打鼾 | 4 | 192 | 21 | 213 |
| 每晚打鼾 | 5 | 224 | 30 | 254 |

资料来源: London;Chapman and Hall,1994.

【例 24-2】 某研究者研究减肥药种类、剂量与减肥效果的关系。研究者将 A、B、C 三种减肥药分别给受试者服用,期间服用减肥药的浓度、服用周数、观测人数及有效人数各不相同。实验结果见表 24-2。请问:能否根据表 24-2 资料建立减肥效果与减肥药种类、服用剂量、服用周数之间的定量关系?如何建立?

表 24-2 减肥效果与减肥药种类、服用剂量及服用周数的关系

| 编 号 | 减肥药 | 剂量 (mg/d) | 服用周数 | 观测人数 | 有效人数 |
|-----|-----|-----------|------|------|------|
| 1 | A | 10.23 | 4 | 50 | 44 |
| 2 | A | 7.76 | 3 | 49 | 42 |
| 3 | A | 5.13 | 3 | 46 | 24 |
| 4 | A | 3.80 | 2 | 48 | 16 |
| 5 | A | 2.57 | 1 | 50 | 6 |
| 6 | B | 50.12 | 5 | 48 | 48 |
| 7 | B | 40.74 | 5 | 50 | 47 |
| 8 | B | 30.20 | 5 | 49 | 47 |
| 9 | B | 20.42 | 2 | 48 | 34 |
| 10 | B | 10.00 | 1 | 48 | 18 |
| 11 | C | 25.12 | 5 | 50 | 48 |
| 12 | C | 20.42 | 4 | 46 | 43 |
| 13 | C | 15.14 | 3 | 48 | 38 |
| 14 | C | 10.00 | 2 | 46 | 27 |

注：“有效人数”是指服药后体重下降至少达到 5 公斤的人数。

24.1.2 对数据结构的分析

例 24-1 中的数据，受试对象为符合入选条件（即满足纳入标准且不满足排除标准）的人，观测并记录其打鼾及罹患心脏病的情况，其中打鼾情况表示原因，为自变量；罹患心脏病的情况表示结果，为二值的响应变量。

例 24-2 中的数据，受试对象为符合入选条件（即满足纳入标准且不满足排除标准）的减肥者，观测内容为服用减肥药的种类、剂量、服用时间及减肥效果，真正的响应变量为“减肥效果”；而减肥药的种类、剂量、服用时间均表示原因，为自变量。

24.1.3 分析目的与统计分析方法的选择

对于例 24-1 和例 24-2 中的资料，统计学分析的目的都是为了建立观测个体产生某种响应的概率与各自变量水平的关系，以便通过某观测个体各自变量的水平取值来预测其产生某种响应的概率。这可以通过 logistic 回归分析来间接实现，也可以通过 Probit 回归分析来直接实现。Probit 回归分析与 logistic 回归分析最大的不同点是：Probit 回归分析中响应变量不再是二值变量（取值 0 或 1，如是否罹患心脏病），而是介于 0 ~ 1 之间的百分比变量（即某事件发生的概率）。这里，采用 Probit 回归分析对资料进行处理。

24.2 Probit 回归分析应用场合及关键技术

24.2.1 适于进行 Probit 回归分析的资料表达形式

Probit 回归分析资料的表达形式如表 24-3 所示。

表 24-3 从实际问题中归纳出的数据结构

| m 个自变量的 水平组合 编号 | 自变量 1 (X_1) | 自变量 2 (X_2) | ... | 自变量 m (X_m) | 观测个体数 (n) | 有某种响应的观测个体数 (resp) |
|-----------------------|--------------------|--------------------|-----|--------------------|--------------|-----------------------|
| 1 | x_{11} | x_{21} | ... | x_{m1} | n_1 | $resp_1$ |
| ... | ... | ... | ... | ... | ... | ... |
| k | x_{1k} | x_{2k} | ... | x_{mk} | n_k | $resp_k$ |

注：m 个自变量的水平组合共有 k 种不同的情况，但每个自变量的水平数并不等于 k，每个自变量在行上的 k 个水平取值中，至少有 2 个水平取值不同即可。

24.2.2 Probit 回归模型简介

logistic 回归模型与 probit 回归模型是现代统计学中对属性变量及其观测数据进行分析的两个很重要的模型，而且，两个模型都是用于在回归分析的范围内分析定性(二值或多值的)响应变量与自变量的关系。probit 回归分析与 logistic 回归分析最大的不同点是：probit 回归分析中响应变量不再是二值变量(取值 0 或 1，如是否罹患心脏病)，而是介于 0 ~ 1 之间的百分比变量(如 resp/n)。probit 回归分析中采用的关联函数为 probit 函数，响应变量取值为 1 的概率为：

$$\begin{aligned} P(y = 1 \mid x) &= \Phi(\beta_0 + \beta_1 x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_0 + \beta_1 x} e^{-\frac{t^2}{2}} dt \\ &= \text{PROBNORM}(\beta_0 + \beta_1 x) \end{aligned} \tag{24-1}$$

这个函数实际上就是标准正态分布下分位数函数 $u = \text{PROBIT}(P)$ 的反函数。式(24 - 1)中的 $\beta_0 + \beta_1 x$ 就相当于这里的 u 。正因为如此，所以 probit 回归分析又称为概率单位回归分析。

24.3 Probit 回归分析的 SAS 实现

24.3.1 对例 24-1 资料进行 Probit 回归分析

1. SAS 程序及说明

所编写的 SAS 程序如下，程序名为：SASTJFX24_1.SAS。

```
DATA SNORE;
INPUT SNORING DISEASE TOTAL @@ ;
CARDS;
0 24 1379 2 35 638 4 21 213
5 30 254
;
RUN;
ods html;
PROC PROBIT DATA = SNORE;

MODEL DISEASE/TOTAL = SNORING/
ITPRINT LACKFIT;
OUTPUT OUT = b p = prob std = std
xbeta = xbeta;
RUN;
proc print data = b;
run;
ods html close;
QUIT;
```

【程序说明】 首先建立数据集 SNORE，输入数据，变量 SNORING、DISEASE、TOTAL 分别表示打鼾情形的打分、每种打鼾情形的罹患心脏病的人数、每种打鼾情形的观察人数。第二部分调用 PROBIT 回归分析过程，MODEL 语句中的“DISEASE/TOTAL =”使用的是事件/试验(events/trials)模型句法，这种句法仅可应用于二值响应数据，“ITPRINT”打印迭代过程、最后计算梯度和二阶微商阵(Hessian)，“LACKFIT”执行 Pearson 卡方拟合检验和对数似然比卡方拟合检验。OUTPUT 语句生成一个新的 SAS 数据集，包括输入数据集中的所有变量、拟合的概率、 $x'\beta$ 的估计及其标准差的估计。“p = prob std = std xbeta = xbeta”规定包含在输出数据集 b 中的统计量，prob 为累计概率的估计，std 为 $a_j + x'b$ 的标准误估计，xbeta 为 $a_j + x'\beta$ 的估计。

2. 主要分析结果及解释

Iteration History for Parameter Estimates

Iter Ridge Loglikelihood Intercept SNORING

| | | | | |
|-----|-----|------------|--------------|--------------|
| 0 | 0 | -1721.7776 | 0 | 0 |
| 1 | 0 | -537.92707 | -1.211021484 | 0.0502278276 |
| ... | ... | ... | ... | ... |
| 6 | 0 | -418.39714 | -1.06055093 | 0.187770292 |

以上是 Probit 过程输出的迭代过程, 对每次迭代给出了迭代序号、修正的 Newton-Raphson 最优化过程的岭脊参数(Ridge)、对数似然的当前值、截距参数(INTERCPT)和 SNORING 的当前估计。

可以得知在第 6 次迭代后, 模型中的系数部分会收敛, 得到如下的 Probit 模型:

$$\begin{aligned}
 P(\text{有心脏病} | \text{打鼾分数} = x) &= \Phi(-2.0606 + 0.1878x) \\
 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-2.0606 + 0.1878x} e^{-\frac{t^2}{2}} dt \\
 &= \text{PROBNORM}(-2.0606 + 0.1878x) \quad (24-2)
 \end{aligned}$$

Last Evaluation of the Negative
of the Gradient

| | |
|-------------|--------------|
| Intercept | SNORING |
| 0.000155387 | 0.0000508314 |

Last Evaluation of the Negative of the
Hessian

| | | |
|-----------|--------------|--------------|
| | Intercept | SNORING |
| Intercept | 470.31291599 | 1060.0837032 |
| SNORING | 1060.0837032 | 4180.3274807 |

Algorithm converged.

以上给出的是最后迭代时的梯度和 Hessian 阵。

Goodness-of-Fit Tests

| Statistic | Value | DF | Pr > ChiSq |
|--------------------|--------|----|------------|
| Pearson Chi-Square | 1.9101 | 2 | 0.3848 |
| L. R. Chi-Square | 1.8716 | 2 | 0.3923 |

以上给出的是由选项 LACKFIT 要求的 Pearson 模型拟合优度检验的卡方值(1.9101)和基于似然比的模型拟合优度检验卡方值(1.8716)及相应的 P 值。由于 P 值分别为 0.3848 和 0.3923, 均大于 0.05, 表明模型是合适的。

其中, 用 Pearson 的卡方检验来检验模型的拟合优度时, 检验假设为:

$$H_0: P(\text{有心脏病} | X = x) = \Phi(\beta_0 + \beta_1 x) \quad \text{for some } \beta_0, \beta_1$$

$$H_1: P(\text{有心脏病} | X = x) \neq \Phi(\beta_0 + \beta_1 x) \quad \text{for any } \beta_0, \beta_1$$

Since the chi-square is small ($p > 0.1000$), fiducial limits will be calculated using a t value of 1.96.

Type III Analysis of Effects

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|---------|----|-----------------|------------|
| SNORING | 1 | 63.1431 | < .0001 |

Analysis of Parameter Estimates

| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | Chi-Square | Pr > ChiSq |
|-----------|----|----------|----------------|-----------------------|------------|------------|
|-----------|----|----------|----------------|-----------------------|------------|------------|

| | | | | | | | |
|-----------|---|---------|--------|---------|---------|--------|--------|
| Intercept | 1 | -1.0606 | 0.0704 | -1.1986 | -1.9225 | 855.49 | <.0001 |
| SNORING | 1 | 0.1878 | 0.0236 | 0.1415 | 0.2341 | 63.14 | <.0001 |

以上给出的是参数估计值及其标准误、检验该参数是否为 0 的近似卡方检验统计量和相应的概率。由参数的估计可得出：

$$P = (\text{有心脏病} | \text{打鼾分数} = x) = \Phi(b_0 + b_1x) = \Phi(-2.0606 + 0.1878x) \quad (24-3)$$
此式与式(24-1)一致。

| | |
|--|-----------|
| Probit Model in Terms of Tolerance Distribution | |
| MU | SIGMA |
| 10.9737856 | 5.3256561 |

这里， $\mu = 10.9737856$ 是刺激的均值(此处指半数致病分数)，且 $\mu = -b_0/b_1$ ； $\sigma = 5.3256561$ 是刺激的尺度参数，且 $\sigma = 1/b_1$ 。

| | | |
|---|----------|----------|
| Estimated Covariance Matrix for Tolerance Parameters | | |
| | MU | SIGMA |
| MU | 1.264472 | 0.735450 |
| SIGMA | 0.735450 | 0.449180 |

这是参数 μ 、 σ 的协方差矩阵。

| | | | | | | |
|-----|---------|---------|-------|---------|----------|----------|
| Obs | SNORING | DISEASE | TOTAL | prob | xbeta | std |
| 1 | 0 | 24 | 1379 | 0.01967 | -1.06055 | 0.070449 |
| 2 | 2 | 35 | 638 | 0.04599 | -1.68501 | 0.046500 |
| 3 | 4 | 21 | 213 | 0.09519 | -1.30947 | 0.061875 |
| 4 | 5 | 30 | 254 | 0.13100 | -1.12170 | 0.079603 |

以上是打印输出的原始数据及根据所拟合的 Probit 回归模型(式(24-3))计算得到的概率预测值(prob 列)、 $(\beta_0 + \beta_1x)$ 的估计值(xbeta 列)及其标准误的估计值(std 列)。

24.3.2 对例 24-2 资料进行 Probit 回归分析

1. SAS 程序及说明

所编写的 SAS 程序如下，程序名为 SASTJFX24_2.SAS。

```
DATA BEAUTY;
INFILE "D:\SASTJFX\SASTJFX24_2.DAT";
INPUT MEDICINE $ DOSE WEEK n RESP @@ ;
LDOSE = LOG10 (DOSE) ;
A=0;B=0;
IF MEDICINE = 'A' THEN A=1;
IF MEDICINE = 'B' THEN B=1;
RUN;
ods html;
proc probit data =BEAUTY;
CLASS MEDICINE;
model resp/n = MEDICINE LDOSE WEEK/
lackfit itprint;

model resp/n=MEDICINE
LDOSE/lackfit itprint;
output out =b p=prob std =std
xbeta =xbeta;
run;
proc probit data =BEAUTY;
model resp/n =A B LDOSE
WEEK/lackfit itprint;
output out =b p=prob std =std
xbeta =xbeta;
run;
proc probit data =BEAUTY;
model resp/n =A B LDOSE/lackfit
```

```

output out = b p = prob
std = std xbeta = xbeta;
run;
proc probit data = BEAUTY;
CLASS MEDICINE;

```

```

itprint;
output out = b p = prob std = std
xbeta = xbeta;
run;
ods html close;
quit;

```

【程序说明】 数据步中，“LDOSE = LOG10(DOSE)”是将剂量取常用对数替换原来的数据，“A = 0; B = 0; IF MEDICINE = 'A' THEN A = 1; IF MEDICINE = 'B' THEN B = 1;”给药物种类赋哑变量。数据步后面写了 4 个 Probit 过程步，其中的第一个过程步，模型中的自变量包括定性变量 MEDICINE 和定量变量 LDOSE 与 WEEK，故用 CLASS 语句在 MODEL 语句之前对定性变量 MEDICINE 进行了说明。由于执行完第一个过程步之后发现变量 WEEK 所对应的系数与 0 的差别无统计学意义，故将该变量从第一个过程步的模型中去掉便得到第二个过程步；第三个过程步和第四个过程步分别与第一个过程步和第二个过程步相对应，只不过第三个过程步和第四个过程步使用了哑变量 A 和 B 来代替定性变量 MEDICINE，因此，MODEL 语句中不再出现变量 MEDICINE，过程步中也不再需要 CLASS 语句。

2. 主要输出结果及其解释

| Probit Procedure | | | | | | | |
|---|-------|---------------|--------------|--------------|--------------|--------------|--------------|
| Iteration History for Parameter Estimates | | | | | | | |
| Iter | Ridge | Loglikelihood | Intercept | MEDICINEA | MEDICINEB | LDOSE | WEEK |
| 0 | 0 | -468.56749 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | -192.4377 | -1.769805965 | 0.808833038 | -0.657091154 | 3.090533775 | -0.060667894 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 6 | 0 | -178.64218 | -2.510779851 | 0.9210000523 | -0.59262653 | 3.5894637343 | 0.0548872216 |

以上是 Probit 过程输出的迭代过程，对每次迭代给出了迭代序号，修正的 Newton - Raphson 最优化过程的岭脊参数 (Ridge)，对数似然的当前值，截距参数 (INTERCPT)、MEDICINEA、MEDICINEB、LDOSE 和 WEEK 的当前估计。

可以得知在第 6 次迭代后，模型中的系数部分会收敛，得到如下的 Probit 模型：

$$P(\text{减肥有效} | \text{MEDICINE, LDOSE, WEEK}) = \Phi(-3.511 + 0.921 * \text{MEDICINEA} - 0.593 * \text{MEDICINEB} + 3.589 * \text{LDOSE} + 0.055 * \text{WEEK}) \quad (24.4)$$

| Class Level Information | | | |
|-------------------------|-----------|----------|------------|
| Name | Levels | Values | |
| MEDICINE | 3 | A | B C |
| Parameter Information | | | |
| Parameter | Effect | MEDICINE | |
| Intercept | Intercept | | |
| MEDICINEA | MEDICINE | A | |
| MEDICINEB | MEDICINE | B | |
| MEDICINEC | MEDICINE | C | |
| LDOSE | LDOSE | | |
| WEEK | WEEK | | |
| Goodness-of-Fit Tests | | | |
| Statistic | Value | DF | Pr > ChiSq |

| | | | |
|--------------------|--------|---|--------|
| Pearson Chi-Square | 7.4265 | 9 | 0.5928 |
| L. R. Chi-Square | 7.5794 | 9 | 0.5770 |

这是由选项 LACKFIT 要求的 Pearson 模型拟合优度检验的卡方值(7.4265)和基于似然比的模型拟合优度检验卡方值(7.5794)及相应的 P 值。由于 P 值分别为 0.5928 和 0.5770, 均大于 0.05, 表明模型是合适的。

Response-Covariate Profile

| | |
|----------------------------|----|
| Response Levels | 2 |
| Number of Covariate Values | 14 |

Since the chi-square is small($p > 0.1000$), fiducial limits will be calculated using a t value of 1.96.

Type III Analysis of Effects

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|----------|----|-----------------|------------|
| MEDICINE | 2 | 10.0777 | 0.0065 |
| LDOSE | 1 | 19.5742 | <.0001 |
| WEEK | 1 | 0.1768 | 0.6742 |

Analysis of Parameter Estimates

| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | Chi-Square | Pr > ChiSq |
|------------|----|----------|----------------|-----------------------|------------|------------|
| Intercept | 1 | -2.5108 | 0.6052 | -4.6969 -1.3247 | 33.65 | <.0001 |
| MEDICINE A | 1 | 0.9210 | 0.3444 | 0.2460 1.5960 | 7.15 | 0.0075 |
| MEDICINE B | 1 | -0.5926 | 0.2208 | -1.0254 -0.1599 | 7.20 | 0.0073 |
| MEDICINE C | 0 | 0.0000 | 0.0000 | 0.0000 0.0000 | . | . |
| LDOSE | 1 | 3.5895 | 0.8113 | 1.9993 5.1796 | 19.57 | <.0001 |
| WEEK | 1 | 0.0549 | 0.1305 | -0.2010 0.3107 | 0.18 | 0.6742 |

这里给出的是参数估计值及其标准误、检验该参数是否为 0 的近似卡方检验统计量和相应的概率。可以看出: WEEK 项的系数与 0 的差别无统计学意义($P = 0.6742 \gg 0.05$), 而其余各项系数与 0 的差别均有统计学意义(P 值均小于 0.01)。

由于“WEEK”项的系数与 0 的差别无统计学意义, 故将其从模型中去掉, 以便提高对误差的估计精度, 更好地拟合模型。本例通过第二个 Probit 过程步来实现。

以下是第二个 Probit 过程的输出结果, 参照对第一个 Probit 过程输出结果的解释进行解释即可, 部分结果不再赘述。

Probit Procedure

Iteration History for Parameter Estimates

| Iter | Ridge | Loglikelihood | Intercept | MEDICINEA | MEDICINEB | LDOSE |
|------|-------|---------------|--------------|--------------|--------------|--------------|
| 0 | 0 | -468.56749 | 0 | 0 | 0 | 0 |
| 1 | 0 | -192.42732 | -1.541584055 | 0.6829667439 | -0.591917225 | 2.7298373652 |
| ... | ... | ... | ... | ... | ... | ... |
| 6 | 0 | -178.73114 | -2.711482371 | 1.0385200758 | -0.654062597 | 3.9061348645 |

Class Level Information

| Name | Levels | Values |
|----------|--------|--------|
| MEDICINE | 3 | A B C |

Parameter Information

| Parameter | Effect | MEDICINE |
|-----------|--------|----------|
|-----------|--------|----------|

| | | |
|-----------|-----------|---|
| Intercept | Intercept | |
| MEDICINEA | MEDICINE | A |
| MEDICINEB | MEDICINE | B |
| MEDICINEC | MEDICINE | C |
| LDOSE | LDOSE | |

Algorithm converged.

Goodness-of-Fit Tests

| Statistic | Value | DF | Pr > ChiSq |
|--------------------|--------|----|------------|
| Pearson Chi-Square | 7.4775 | 10 | 0.6797 |
| L. R. Chi-Square | 7.7573 | 10 | 0.6525 |

Response-Covariate Profile

| | |
|----------------------------|----|
| Response Levels | 2 |
| Number of Covariate Values | 14 |

Since the chi-square is small($p > 0.1000$), fiducial limits will be calculated using a t value of 1.96

Type III Analysis of Effects

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|----------|----|-----------------|------------|
| MEDICINE | 2 | 55.9682 | <.0001 |
| LDOSE | 1 | 162.0068 | <.0001 |

Analysis of Parameter Estimates

| Parameter | | DF | Estimate | Standard Error | 95% Confidence Limits | Chi-Square | Pr > ChiSq |
|-----------|---|----|----------|----------------|-----------------------|------------|------------|
| Intercept | | 1 | -2.7115 | 0.3749 | -4.4463 -1.9767 | 98.01 | <.0001 |
| MEDICINE | A | 1 | 1.0385 | 0.2029 | 0.6409 1.4362 | 26.20 | <.0001 |
| MEDICINE | B | 1 | -0.6541 | 0.1645 | -0.9764 -0.3317 | 15.82 | <.0001 |
| MEDICINE | C | 0 | 0.0000 | 0.0000 | 0.0000 0.0000 | . | . |
| LDOSE | | 1 | 3.9061 | 0.3069 | 3.3046 4.5076 | 162.01 | <.0001 |

从模型拟合优度检验的结果看,第二个 Probit 模型比第一个 Probit 模型拟合得好。根据第二个 Probit 过程的输出结果,得到最终的 Probit 回归方程为:

$$P(\text{减肥有效} | \text{MEDICINE}, \text{LDOSE}, \text{WEEK}) = \Phi(-3.7115 + 1.0385 * \text{MEDICINEA} - 0.6541 * \text{MEDICINEB} + 3.9061 * \text{LDOSE}) \quad (24-5)$$

以下是第三个 Probit 过程步的输出结果(与第一个 Probit 过程的输出结果在实质上是相同的):

Probit Procedure

Iteration History for Parameter Estimates

| Iter | Ridge | Loglikelihood | Intercept | A | B | LDOSE | WEEK |
|------|-------|---------------|--------------|--------------|--------------|--------------|--------------|
| 0 | 0 | -468.56749 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | -192.4377 | -1.769805965 | 0.808833038 | -0.657091154 | 3.090533775 | -0.060667894 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 6 | 0 | -178.64218 | -2.510779851 | 0.9210000523 | -0.59262653 | 3.5894637343 | 0.0548872216 |

Algorithm converged.

Goodness-of-Fit Tests

| Statistic | Value | DF | Pr > ChiSq |
|-----------|-------|----|------------|
|-----------|-------|----|------------|

| | | | |
|--------------------|--------|---|--------|
| Pearson Chi-Square | 7.4265 | 9 | 0.5928 |
| L. R. Chi-Square | 7.5794 | 9 | 0.5770 |

Response-Covariate Profile

| | |
|----------------------------|----|
| Response Levels | 2 |
| Number of Covariate Values | 14 |

Since the chi-square is small($p > 0.1000$), fiducial limits will be calculated using a t value of 1.96.

Type III Analysis of Effects

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|--------|----|-----------------|------------|
| A | 1 | 7.1512 | 0.0075 |
| B | 1 | 7.2050 | 0.0073 |
| LDOSE | 1 | 19.5742 | <.0001 |
| WEEK | 1 | 0.1768 | 0.6742 |

Analysis of Parameter Estimates

| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|-----------|----|----------|----------------|-----------------------|---------|------------|------------|
| Intercept | 1 | -2.5108 | 0.6052 | -4.6969 | -1.3247 | 33.65 | <.0001 |
| A | 1 | 0.9210 | 0.3444 | 0.2460 | 1.5960 | 7.15 | 0.0075 |
| B | 1 | -0.5926 | 0.2208 | -1.0254 | -0.1599 | 7.20 | 0.0073 |
| LDOSE | 1 | 3.5895 | 0.8113 | 1.9993 | 5.1796 | 19.57 | <.0001 |
| WEEK | 1 | 0.0549 | 0.1305 | -0.2010 | 0.3107 | 0.18 | 0.6742 |

以下是第四个 Probit 过程输出的结果(与第二个 Probit 过程的输出结果在实质上是相同的):

Probit Procedure

Iteration History for Parameter Estimates

| Iter | Ridge | Loglikelihood | Intercept | A | B | LDOSE |
|------|-------|---------------|--------------|--------------|--------------|--------------|
| 0 | 0 | -468.56749 | 0 | 0 | 0 | 0 |
| 1 | 0 | -192.42732 | -1.541584055 | 0.6829667439 | -0.591917225 | 2.7298373652 |
| ... | ... | ... | ... | ... | ... | ... |
| 6 | 0 | -178.73114 | -2.711482371 | 1.0385200758 | -0.654062597 | 3.9061348645 |

Model Information

| | |
|------------------------|--------------|
| Data Set | WORK.BEAUTY |
| Events Variable | RESP |
| Trials Variable | n |
| Number of Observations | 14 |
| Number of Events | 482 |
| Number of Trials | 676 |
| Name of Distribution | Normal |
| Log Likelihood | -178.7311379 |

Last Evaluation of the Negative of the Gradient

| Intercept | A | B | LDOSE |
|--------------|--------------|--------------|--------------|
| -1.680064E-9 | -8.66347E-11 | -1.330463E-9 | -4.217919E-9 |

Last Evaluation of the Negative of the Hessian

| | Intercept | A | B | LDOSE |
|-----------|--------------|--------------|--------------|--------------|
| Intercept | 274.36023311 | 119.30764193 | 80.103969139 | 273.76803782 |
| A | 119.30764193 | 119.30764193 | -7.10543E-15 | 84.452626278 |

| | | | | |
|-------|--------------|--------------|--------------|--------------|
| B | 80.103969139 | -7.10543E-15 | 80.103969139 | 102.21002082 |
| LDOSE | 273.76803782 | 84.452626278 | 102.21002082 | 302.04881847 |

Algorithm converged.

Goodness-of-Fit Tests

| Statistic | Value | DF | Pr > ChiSq |
|--------------------|--------|----|------------|
| Pearson Chi-Square | 7.4775 | 10 | 0.6797 |
| L. R. Chi-Square | 7.7573 | 10 | 0.6525 |

Response-Covariate Profile

| | |
|----------------------------|----|
| Response Levels | 2 |
| Number of Covariate Values | 14 |

Since the chi-square is small ($p > 0.1000$), fiducial limits will be calculated using a t value of 1.96.

Type III Analysis of Effects

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|--------|----|-----------------|------------|
| A | 1 | 26.1995 | <.0001 |
| B | 1 | 15.8179 | <.0001 |
| LDOSE | 1 | 162.0068 | <.0001 |

Analysis of Parameter Estimates

| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | Chi-Square | Pr > ChiSq |
|-----------|----|----------|----------------|-----------------------|------------|------------|
| Intercept | 1 | -2.7115 | 0.3749 | -4.4463 -1.9767 | 98.01 | <.0001 |
| A | 1 | 1.0385 | 0.2029 | 0.6409 1.4362 | 26.20 | <.0001 |
| B | 1 | -0.6541 | 0.1645 | -0.9764 -0.3317 | 15.82 | <.0001 |
| LDOSE | 1 | 3.9061 | 0.3069 | 3.3046 4.5076 | 162.01 | <.0001 |

24.4 PROBIT 过程语法简介

PROBIT 过程可以包括以下语句：

```
PROC PROBIT <options>;
MODEL response = independents </options>;
BY variables;
CLASS variables;
OUTPUT <OUT = SAS-data-set> <options>;
WEIGHT variable;
CDFPLOT <VAR = variable> <options>;
INSET <keyword - list> </options>;
IPPPLOT <VAR = variable> <options>;
LPREDPLOT <VAR = variable> <options>;
PREDPPLOT <VAR = variable> <options>;
```

主要语句的说明如下。

1. PROC PROBIT 语句

调用 PROBIT 过程，可用的选项如下。

(1) 数据集选项

DATA = SAS-data-set——给出被 PROBIT 过程使用的 SAS 数据集的名字，若省略该选项，则使用最近生成的 SAS 数据集。

OUTEST = SAS-data-set——生成一个输出数据集，包含参数估计及参数估计的协方差阵（可选择的）。若未规定此选项，则不创建输出数据集。

COVOUT——把估计的协方差写入 OUTEST = 的数据集中。

(2) 模型选项

C = rate|OPTC——控制如何处理自然响应。规定 OPTC 时要求自然（阈值）响应比率 C 的估计被优化；规定 C = rate 时给出自然响应比率或自然响应比率的初始估计。该值必须在 0 和 1 之间。

ORDER = FREQ|DATA|INTERNAL|FORMATTED——规定分类变量（在 CLASS 语句中规定的）水平和排列次序，这个次序确定数据中的每个水平对应于模型中的哪个参数。缺省时 ORDER = FORMATTED。

LOG|LN——把第一个连续的自变量用它的自然对数替换后分析数据。

LOG10——规定的分析像 LOG 或 LN 一样，不同的是用常用对数（以 10 为底的对数）而不是自然对数。

INVERSECT——对获得选择响应值的第一个连续的自变量（通常是剂量）的值计算置信界限。

(3) 模型的拟合检验选项

LACKFIT——执行 Pearson 卡方拟合检验和对数似然比卡方拟合检验。如果在 PROBIT 过程运行之前进行拟合检验，数据集必须按自变量排序。

HPROB = p——规定一个 0.10 以外的概率水平来给出拟合的优度。

(4) 打印输出选项

NOPRINT——不打印输出。

2. MODEL 语句

有以下两种写法：

```
<label>:MODEL response=independents</options>;
<label>:MODEL events/trials=independents</options>;
```

MODEL 语句用来指明响应变量和自变量。还可用来指明响应变量的分布及其他选项。通常，SAS 软件不支持在一个 PROBIT 过程中规定多个 MODEL 语句。若有需要，可以写两个或两个以上的 PROBIT 过程。可选择的 label 用于对来自相应 MODEL 语句的输出加上标签。

(1) 模型说明项

response——单个响应变量，其值用来表示观测的响应水平。这个响应变量必须在 CLASS 语句中列出。

events/trials——命名一对非负数值变量，使得 $0 \leq \text{events/trials} \leq 1$ 。

independents——给出自变量的名字。

(2) 选项(options)

其中用于规定模型的选项有：

D = distribution-type——规定用于模型响应概率的累计分布函数。

INTERCEPT = value——初始化截距参数为 value，缺省时为 0。

INITIAL = values——设置模型里其他参数的初始值。这些值必须按 MODEL 语句中列出的变量顺序给出。如果在 MODEL 语句中列出的自变量有些是分类变量，那么必须对分类变量给

出值的个数为分类水平数减 1 个。选项 INITIAL = 可以指定为一个标准值列表, 该标准值列表可以是下面的任何一种形式:

Initial = 3 4 5 (列表用空格隔开)

Initial = 3, 4, 5 (列表用逗号隔开)

Initial = 3 to 5 (列表采用 x to y 的形式)

Initial = 3 to 5 by 1 (列表采用 x to y by z 的形式)

Initial = 1, 3 to 5, 9 (列表采用以上几种形式的组合形式)

缺省时所有参数的初始值皆被置为 0。

NOINT——拟合没有截距参数的模型。当响应变量为二值变量时, 这个选项很有用。选项 NOINT 设置截距参数对应于最低的响应水平等于 0。

用于拟合模型的选项有:

HPROB = value——规定一个 0.10 以外的概率水平来表示拟合优度。若该选项在 PROC PROBIT 和 MODEL 两个语句中都规定了, 则 MODEL 语句的这个选项优先。

MAXITER = value——为了进行参数估计, 设置最大迭代次数。缺省时为 50。

CONVERGE = value——给出收敛准则。缺省时为 0.001。

SINGULAR = value——在一组自变量里用于确定线性相关的显著性准则。缺省时为 1E-12。

控制打印输出:

CORRB——打印参数估计的相关阵的估计。

COVB——打印参数估计的协方差阵的估计。

INVERSECT——对获得选择响应值的第一个连续的自变量(通常是剂量)的值计算置信界限。当算法不收敛时, 置信限为缺失值。若该选项在 PROC PROBIT 和 MODEL 两个语句中都规定了, 则 MODEL 语句的这个选项优先。

ITPRINT——打印迭代过程, 最后计算梯度和二阶微商阵(Hessian)。

LACKFIT——执行 Pearson 卡方拟合检验和对数似然比卡方拟合检验。如果在 PROBIT 过程运行之前进行拟合检验, 数据集必须按自变量排序。

3. OUTPUT 语句

格式如下:

```
OUTPUT <OUT = SAS-data-set> <keyword -1 = name...keyword -n = name>;
```

OUTPUT 语句生成一个新的 SAS 数据集, 包括输入数据集中的所有变量, 拟合的概率, $x'\beta$ 的估计及其标准差的估计。其中,

SAS-data-set——给出包括这些统计量的输出数据集的名字。若缺省, 新数据集按习惯命名为 DATAn。

keyword = name——规定包含在这个输出数据集中的统计量, 并对包含统计量的新变量给出 name(名字)。有效的 keyword(关键词)为: PROBLP(累计概率的估计), STD(估计量 $a_j + x'b$ 的标准差), XBETA($a_j + x'\beta$ 的估计)。

4. CLASS 语句

用于指定分类变量。如果在 MODEL 语句中给出了单个响应变量, 它也必须在 CLASS 语句中列出。

分类水平由分类变量的格式化值确定。因此可以使用格式把变量的取值分组作为几个水平。如果使用 CLASS 语句，它必须放在 MODEL 语句之前。

5. BY 语句

BY 语句可以和 PROBIT 过程一起使用，以便得到由 BY 变量定义的分组中观测的单独分析。当 BY 语句出现时，过程要求 DATA = 的输入数据集按照 BY 语句中变量的顺序排序。

6. WEIGHT 语句

WEIGHT 语句和 PROBIT 过程一起使用，以便用 WEIGHT 变量规定的值对每个观测加权重。每个观测对似然函数的贡献也应乘以权重变量的值。权重值为 0、负数或缺失值的观测在模型估计中不使用。

24.5 本章小结

本章介绍了负二项回归与 Probit 回归的用法及其 SAS 实现方法。属于广义线性模型的负二项回归模型是基于负二项分布的，理解这个模型的关键在于掌握它与 Poisson 回归模型的联系。负二项回归可以解决 Poisson 回归中存在的过离散问题，将方差表示为均数的二次函数的形式。在 SAS 软件中，GENMOD 过程与 COUNTREG 过程均可用于实现负二项回归分析。而 Probit 回归与 logistic 回归均可用于响应变量为定性变量的回归分析，且两者用于处理同一问题时，效果均很好，而且伯仲难分，至于该用何种模型，则取决于实际问题的具体要求及个人喜好。

(柳伟伟 周诗国 陶丽新)

第 25 章 生存资料 Cox 模型回归分析

在分析多个因素对生存时间的影响时，人们通常希望像一般的回归分析一样，能建立生存时间(因变量或反应变量)随危险因素(自变量或协变量)变化的回归方程，以便对危险因素的作用大小有一个全面的了解和掌握，并根据危险因素的不同取值对生存率(或危险率)进行预测。能实现此目的的生存分析方法有 Cox 模型回归分析和参数模型回归分析。

本章主要介绍 Cox 比例风险回归模型、SAS 软件中的 PHREG 过程及其应用。

25.1 实 例

25.1.1 问题与数据

【例 25-1】 30 例膀胱肿瘤患者生存资料原始记录见表 25-1。研究者欲分析影响膀胱肿瘤患者生存时间长短的因素，包括年龄、肿瘤分级、肿瘤大小和是否复发，并根据影响因素的取值对不同患者的生存情况进行预测。

表 25-1 30 例膀胱肿瘤患者生存资料原始记录

| id | age | grade | size | relapse | start | end | t | status |
|-----|-----|-------|------|---------|------------|------------|-----|--------|
| 1 | 62 | 1 | 0 | 0 | 02/10/1996 | 12/30/2000 | 59 | 0 |
| 2 | 64 | 1 | 0 | 0 | 03/05/1996 | 08/12/2000 | 54 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 30 | 54 | 3 | 1 | 1 | 03/10/2000 | 09/20/2000 | 6 | 1 |

注：age：年龄(岁)；grade：肿瘤分级，I 级 = 1，II 级 = 2，III 级 = 3；size：肿瘤大小(cm)， $\geq 3.0 = 1$ ， $< 3.0 = 0$ ；relapse：是否复发，是 = 1，否 = 0；start：随访起点(月/日/年)；end：随访终点(月/日/年)；t：生存时间(月)；status：生存结局，死亡 = 1，截尾 = 0。

25.1.2 对数据结构的分析

表 25-1 中，以原始数据的形式记录了 30 例膀胱肿瘤患者的有关情况，包括年龄、肿瘤分级、肿瘤大小、是否复发、随访起点、随访终点、生存时间和生存结局等。其中，年龄和生存年月是定量指标，肿瘤分级、肿瘤大小、是否复发和生存结局都是定性指标。由于存在删失数据，此资料为生存资料。当然，该资料涉及多个影响因素。所以，此资料应为多因素设计一元生存资料。

25.1.3 分析目的与统计分析方法的选择

研究者希望分析影响膀胱肿瘤患者生存时间长短的因素，并根据影响因素的取值对不同患者的生存情况进行预测。此时，可根据生存时间的分布情况选择合适的生存分析方法。

当生存时间的准确分布无法获得时，前述目的无法直接实现。此时，可采用 Cox 模型回归分析。此模型在形式上与参数模型相似，但对模型中各参数进行估计时不依赖于特定分布的假设，所以又称半参数模型。

当然，在可以通过图示法或统计检验法，得到待分析的生存资料服从某特定分布的参数模型时，如指数分布模型或 Weibull 分布模型时，可采用生存资料的参数模型回归分析直接拟合这些模型，所得结果将更加准确。

本章主要介绍 Cox 模型回归分析，故对生存时间的分布未加以详查。本章将详细介绍生存资料的参数模型回归分析。

25.2 生存资料 Cox 模型回归分析简介

目前对生存资料的多因素分析最常用的方法是 Cox 比例风险回归模型(proportional hazards regression model)，简称 Cox 模型。该模型是一种多因素的生存分析方法，它可同时分析众多因素对生存期的影响，分析带截尾生存时间的资料，且不要求估计资料的生存分布类型。由于上述优良性质，该模型自英国统计学家 D. R. Cox 于 1972 年提出以来，在医学随访研究中得到非常广泛的应用。

Cox 模型属比例风险模型簇，其基本假定之一是比例风险假定(简称 PH 假定)。只有在满足该假定前提条件下，基于此模型的分析预测才是可靠有效的。正像我们所熟知的 t 检验中的正态分布假定一样，当使用比例风险模型时，比例风险假定应看成一个基本前提。

检查某协变量是否满足 PH 假定，最简单的方法是观察按该变量分组的 Kaplan-Meier 生存曲线，若生存曲线交叉，提示不满足 PH 假定。第二种方法是绘制按该变量分组的 $\ln[-\ln\hat{S}(t)]$ 对生存时间 t 的图，曲线应大致平行或等距。如各协变量均满足或近似满足 PH 假定，可直接应用基本 Cox 模型。

25.3 生存资料 Cox 模型回归分析的 SAS 实现

25.3.1 SAS 程序

表 25-1 数据中，年龄为定量变量，将年龄转化为两分类变量(<60 岁和≥60 岁)，肿瘤分级、肿瘤大小和是否复发均为分类变量，按这四个变量的水平分组，采用 Kaplan-Meier 法绘制生存曲线。除两年龄组生存曲线[图 25-1(a)]在近 30 月处略有交叉外，图 25-1(b)~(d)中曲线均基本无交叉，说明四个变量基本满足 PH 假定。所以，可进行生存资料的 Cox 模型回归分析。

下面借助 SAS 软件实现本生存资料的 Cox 模型回归分析，程序名为 SASTJFX25_1.SAS。

| | |
|---|---|
| <pre>DATA SASTJFX25_1; INPUT age grade size relapse t status @@ ; CARDS; 62 1 0 0 59 0 64 1 0 0 54 1 54 3 1 1 6 1 ; ods html;</pre> | <pre>PROC PHREG; MODEL t* status(0) =age grade size relapse/SELECTION = STEPWISE SLE =0.05 SLS =0.05 RL; OUTPUT OUT =report SURVIVAL = s XBETA =pi/ORDER = DATA METHOD = PL; RUN; PROC PRINT DATA =report; RUN; ods html close;</pre> |
|---|---|

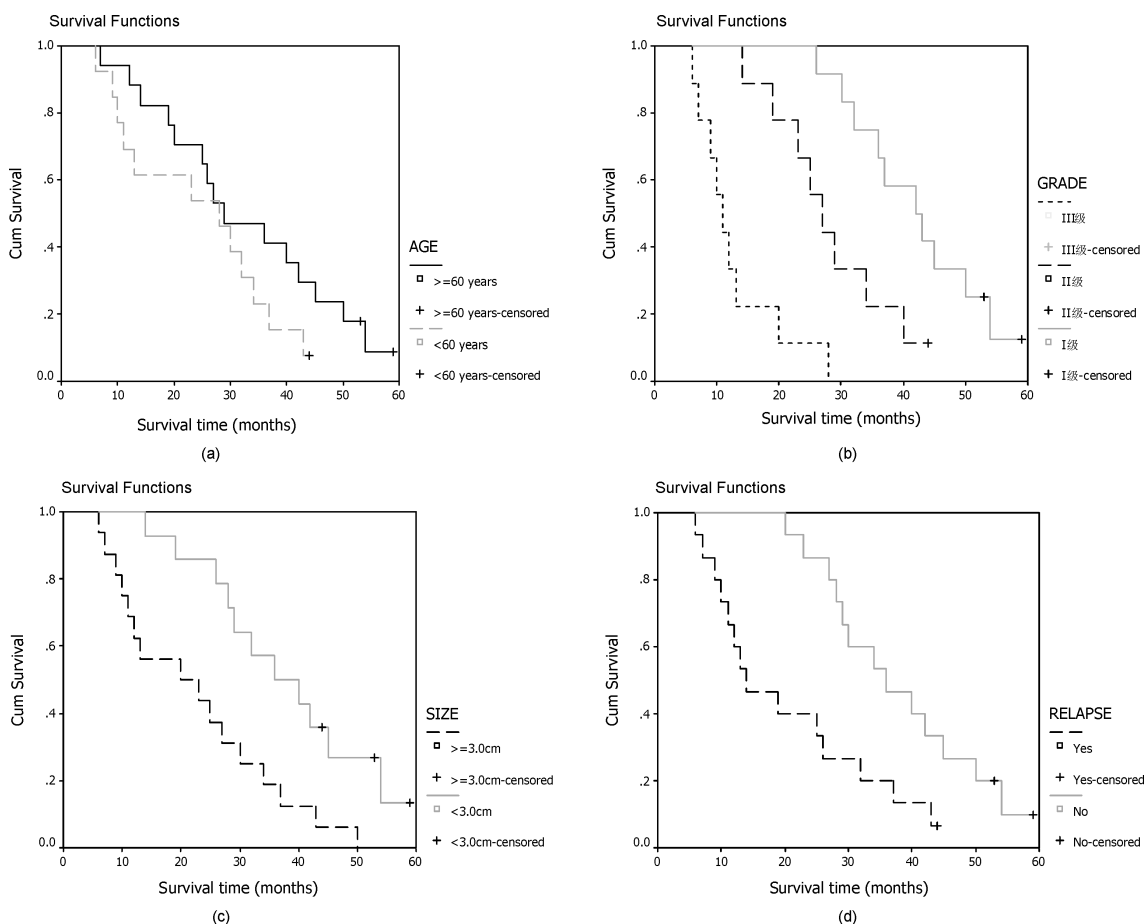


图 25-1 膀胱肿瘤患者数据四个变量的生存曲线
(A 年龄; B 肿瘤分级; C 肿瘤大小; D 是否复发)

【程序说明】 PHREG 过程是实现 Cox 模型回归分析的标准过程，其中 MODEL 语句是必需语句。MODEL 语句左边为生存时间和生存结局变量(括号内为截尾值)，右边为协变量。该语句中比较重要的选项有：

TIES = DISCRETE|EXACT|BRESLOW|EFRON，指定重合生存时间或称结点(ties)的处理方法。DISCRETE 和 EXACT 为精确法，DISCRETE 法假定事件确实发生在相同的时间，而 EXACT 法假定结点来自连续的无结点资料。BRESLOW 法(1974，缺省值)和 EFRON 法(1977)是精确法的近似。关于四种方法的选择，没有结点时，四种方法结果相同；结点比例不是很大时，四种方法结果相近；结点比例很大时，两种近似结果有偏性，考虑计算耗时，可选 EFRON 近似法。

SELECTION = FORWARD|BACKWARD|STEPWISE|NONE|SCORE，指定变量筛选的方法，分别表示前进法、后退法、逐步法、全回归模型(缺省值)和最优子集法。

SLE = 和 SLS = 分别指定引入和剔除变量的显著性水平 α 。缺省值为 $\alpha = 0.05$ 。

RL 要求输出风险比 HR 的 $100(1 - \alpha)\%$ 置信限。

OUTPUT 语句创建一个新的 SAS 数据集，含有为每一个观测计算的一些统计量，SAS 为每一个统计量定义一个关键字，如生存率和预后指数分别用 SURVIVAL 和 XBETA 表示。选项

ORDER = DATA 规定输出的数据集中的观测顺序与输入数据集中的顺序一致；METHOD = PL|CH|EMP 规定用于计算生存率的方法，PL 表示生存率的乘积极限估计法(缺省值)，CH 和 EMP 表示生存率的经验累积风险函数估计法。

PHREG 过程的其他语句有：

(1)STRATA 语句用于建立分层 Cox 模型。格式为：STRATA 分层变量。注意分层变量不宜太多，否则每层观察单位数太少，会影响分析结果。

(2)BASELINE 语句用于输出对原有数据集以外某新数据集中观测的生存预测。格式为：BASELINE OUT = 输出数据集名 COVARIATES = 新数据集名关键字 = 变量名/选项。新数据集中的变量必须与最后 Cox 模型中的变量相对应。常用关键字有 SURVIVAL(生存率)和 XBETA(预后指数)等。该语句中“/”后的选项有 METHOD = PL|CH|EMP，意义同 OUTPUT 语句。

25.3.2 主要分析结果及解释

Model Information

| | |
|--------------------|------------------|
| Data Set | WORK.SASTJFX25_1 |
| Dependent Variable | t |
| Censoring Variable | status |
| Censoring Value(s) | 0 |
| Ties Handling | BRESLOW |

Summary of the Number of Event and Censored Values

| Total | Event | Censored | Percent Censored |
|-------|-------|----------|------------------|
| 30 | 27 | 3 | 10.00 |

Step 1. Variable grade is entered. The model contains the following explanatory variables: grade

Convergence Status

Convergence criterion(GCONV = 1E - 8) satisfied.

Model Fit Statistics

| Criterion | Without | With |
|-----------|------------|------------|
| | Covariates | Covariates |
| - 2 LOG L | 143.536 | 120.374 |
| AIC | 143.536 | 122.374 |
| SBC | 143.536 | 123.670 |

Testing Global Null Hypothesis: BETA = 0

| Test | Chi-Square | DF | Pr > ChiSq |
|------------------|------------|----|------------|
| Likelihood Ratio | 23.1614 | 1 | < .0001 |
| Score | 26.0945 | 1 | < .0001 |
| Wald | 20.9463 | 1 | < .0001 |

Step 2. Variable size is entered. The model contains the following explanatory variables: grade size

Convergence Status

Convergence criterion(GCONV = 1E - 8) satisfied.

Model Fit Statistics

| Criterion | Without | With |
|-----------|------------|------------|
| | Covariates | Covariates |

| | | |
|-----------|---------|---------|
| - 2 LOG L | 143.536 | 113.738 |
| AIC | 143.536 | 117.738 |
| SBC | 143.536 | 120.330 |

Testing Global Null Hypothesis: BETA = 0

| Test | Chi-Square | DF | Pr > ChiSq |
|------------------|------------|----|------------|
| Likelihood Ratio | 29.7973 | 2 | < .0001 |
| Score | 30.7819 | 2 | < .0001 |
| Wald | 22.7987 | 2 | < .0001 |

Step 3. Variable relapse is entered. The model contains the following explanatory variables:
grade size relapse

Convergence Status

Convergence criterion(GCONV = 1E - 8)satisfied.

Model Fit Statistics

| Criterion | Without Covariates | With Covariates |
|-----------|-----------------------|--------------------|
| - 2 LOG L | 143.536 | 109.115 |
| AIC | 143.536 | 115.115 |
| SBC | 143.536 | 119.002 |

Testing Global Null Hypothesis: BETA = 0

| Test | Chi-Square | DF | Pr > ChiSq |
|------------------|------------|----|------------|
| Likelihood Ratio | 34.4210 | 3 | < .0001 |
| Score | 33.9810 | 3 | < .0001 |
| Wald | 23.6573 | 3 | < .0001 |

Note: No(additional) variables met the 0.05 level for entry into the model.

Analysis of Maximum Likelihood Estimates

| Variable | DF | Parameter Estimate | Standard Error | Chi-Square | Pr > ChiSq | Hazard Ratio | 95% Hazard Ratio Confidence Limits |
|----------|----|-----------------------|-------------------|------------|------------|-----------------|--|
| grade | 1 | 1.68027 | 0.38164 | 19.3839 | < .0001 | 5.367 | 2.540 11.339 |
| size | 1 | 1.07809 | 0.46004 | 5.4919 | 0.0191 | 2.939 | 1.193 7.241 |
| relapse | 1 | 0.97892 | 0.46023 | 4.5243 | 0.0334 | 2.662 | 1.080 |

Summary of Stepwise Selection

| Step | Variable Entered | Variable Removed | Number In | Score Chi-Square | Wald Chi-Square | Pr > ChiSq |
|------|---------------------|---------------------|--------------|---------------------|--------------------|------------|
| 1 | grade | | 1 | 26.0945 | . | < .0001 |
| 2 | size | | 2 | 6.8436 | . | 0.0089 |
| 3 | relapse | | 3 | 4.7691 | . | 0.0290 |

逐步法的第一步、第二步和第三步分别引入变量肿瘤分级、肿瘤大小和是否复发, SAS 输出每一步的模型拟合统计量(- 2 LOG L、AIC 和 SBC) 及模型检验结果(似然比检验、Score 检验和 Wald 检验), 之后是模型的最大似然估计, 包括参数估计(Parameter Estimate)、估计值标准误(Standard Error)、Wald χ^2 值(Chi-Square)、P 值(Pr > ChiSq)、风险比 HR 及 HR95% 置信区间(Hazard Ratio Confidence Limits)。

Cox 模型结果显示：肿瘤分级、肿瘤大小和是否复发为膀胱肿瘤患者生存时间长短的重要因素。三个变量的回归系数均为正值，提示高肿瘤分级、肿瘤大于或等于 3.0cm 和复发为膀胱肿瘤患者死亡的条件组合（即与三个因素对应的各自坏水平）。肿瘤大小和是否复发保持不变的情形下，肿瘤分级每提升一级，死亡风险就是前一级的 5.367 倍（ $e^{1.68027}$ ）；肿瘤分级和是否复发保持不变的情形下，肿瘤大于或等于 3.0cm 者死亡风险是肿瘤小于 3.0cm 者死亡风险的 2.939 倍（ $e^{1.07809}$ ）；肿瘤分级和肿瘤大小保持不变的情形下，肿瘤复发者死亡风险是未复发者死亡风险的 2.662 倍（ $e^{0.97892}$ ）。

由 Cox 回归分析结果，得出风险函数的表达式为

$$h(t) = h_0(t) \exp(1.680 \times \text{grade} + 1.078 \times \text{size} + 0.979 \times \text{relapse})$$

此表达式右边指数部分取值越大，则风险函数 $h(t)$ 越大，预后越差，故称为预后指数（prognostic index, PI）。

生存率可由下式估计

$$\hat{S}(t) = [\hat{S}_0(t)]^{\exp(\sum \hat{\beta}_i X_i)}$$

式中 $\hat{S}_0(t)$ 为基准生存率，代表所有自变量取值均为 0 的病人在 t 时刻的生存率。可采用 Breslow 估计，

$$\hat{S}_0(t) = \prod_{t_{(j)} \leq t} \left\{ \exp \left[\frac{-d_j}{\sum_{i \in R_j} \exp \left(\sum \hat{\beta}_i X_i \right)} \right] \right\}$$

式中 \prod 为连乘积符号， t_j 为排序后的完全生存时间， d_j 为 t_j 处死亡数， R_j 为 t_j 处风险集。

| Obs | age | grade | size | relapse | t | status | pi | s |
|-----|-----|-------|------|---------|----|--------|---------|---------|
| 1 | 62 | 1 | 0 | 0 | 59 | 0 | 1.68027 | 0.25588 |
| 2 | 64 | 1 | 0 | 0 | 54 | 1 | 1.68027 | 0.25588 |
| 3 | 52 | 2 | 0 | 1 | 44 | 0 | 4.33945 | 0.01783 |
| 4 | 60 | 1 | 0 | 0 | 53 | 0 | 1.68027 | 0.51176 |
| 5 | 59 | 2 | 1 | 0 | 23 | 1 | 4.43863 | 0.66200 |

输出结果显示了表 25-1 中前 5 例患者的预后指数 PI 及其所对应生存时间的生存率（为节省篇幅，其余患者的情况未列出）。如 1 号患者，肿瘤分级 I 级，肿瘤大小 <3.0cm，未复发，PI = 1.68，59 个月生存率 25.6%；3 号患者，肿瘤分级 II 级，肿瘤 <3.0cm，复发，PI = 4.34，44 个月生存率 1.8%。

25.4 本章小结

(1) 本章主要介绍了生存资料的 Cox 模型回归分析，以实例的形式展示了如何借助 SAS 软件实现 Cox 模型回归分析，并对程序输出的结果给出合理的解释。

(2) Cox 模型属比例风险模型、乘法模型。模型中回归系数 β_j 的统计学意义是，其他变量不变条件下，变量 X_j 每变化一个单位所引起的风险比的自然对数，或使风险函数增至 $\exp(\beta_i)$ 倍。Cox 回归可用于影响因素分析、校正混杂因素后的组间比较以及多变量生存预测。SAS 实现过程为 PHREG。

（余红梅 高辉）

第 26 章 生存资料参数模型回归分析

通过图解法或统计检验法，能得知待分析的生存资料服从某特定分布的参数模型时，直接拟合这些模型，可获得比盲目用其他方法更准确的结果。

本章主要介绍生存资料参数回归模型以及 SAS 软件中的 LIFEREG 过程及其应用。

26.1 实 例

26.1.1 问题与数据

【例 26-1】 原文作者在 17 年里追踪调查了 149 位糖尿病患者，数据见表 26-1。变量及其赋值如下，试进行患者生存时间的影响因素分析并进行生存预测。

结局(status, 1 表示死亡; 0 表示截尾); 生存时间(t, 年); 随访开始时年龄(Age1, 岁); 体重指数(BMI); 诊断出糖尿病时的年龄(Age0, 岁); 吸烟状况(sm, 0 表示不吸烟; 1 表示曾吸烟; 2 表示吸烟); 收缩压(SBP, mmHg); 舒张压(DBP, mmHg); 心电图读数(ECG, 0 表示正常; 1 表示可疑; 2 表示异常); 病人是否有冠心病(CHD, 0 表示无; 1 表示有)。

表 26-1 149 位糖尿病患者的生存资料

| Id | status | t | Age1 | BMI | Age0 | smk | SBP | DBP | ECG | CHD |
|-----|--------|------|------|------|------|-----|-----|-----|-----|-----|
| 1 | 0 | 12.4 | 44 | 34.2 | 41 | 0 | 132 | 96 | 0 | 0 |
| 2 | 0 | 12.4 | 49 | 32.6 | 48 | 2 | 130 | 72 | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 149 | 0 | 10.5 | 49 | 30.8 | 47 | 1 | 146 | 86 | 0 | 0 |

资料来源：E. T. Lee. 生存数据分析的统计方法. 陈家鼎, 戴中维译. 北京：中国统计出版社, 1998：72-76.

26.1.2 对数据结构的分析

表 26-1 中，记录了 149 例糖尿病患者的有关情况，包括结局、生存时间、随访开始时年龄、体重指数等。其中，生存时间、随访开始时年龄、体重指数、诊断出糖尿病时的年龄、收缩压和舒张压是定量指标，吸烟状况、心电图读数、病人是否患有冠心病和结局都是定性指标。由于存在删失数据，此资料为生存资料。当然，资料中涉及多个影响因素。所以，此资料应为多因素设计一元生存资料。

26.1.3 分析目的与统计分析方法的选择

研究者希望分析影响糖尿病患者生存时间的因素，并根据影响因素的取值对不同患者的生存情况进行预测。此时，可根据生存时间的分布情况选择合适的生存分析方法。

当生存时间的准确分布无法获得时，前述目的无法直接实现。此时，可采用 Cox 模型回归分析。此模型在形式上与参数模型相似，但对模型中各参数进行估计时不依赖于特定分布的假设，所以又称半参数模型。

当然，在可以通过图示法或统计检验法，得到待分析的生存资料服从某特定分布的参数模

型时，如指数分布模型或 Weibull 分布模型，可采用生存资料的参数模型回归分析，直接拟合这些模型，所得结果将更加准确。

本章主要介绍参数模型回归分析。

26.2 生存资料参数模型回归分析简介

生存资料参数模型回归分析的一个重要内容是模型拟合或分布拟合。描述生存时间分布的模型通常有指数分布、Weibull 分布、对数正态分布、Gamma 分布等。常见生存时间分布的概率密度函数 $f(t)$ 、生存函数 $S(t)$ 和风险函数 $h(t)$ 见表 26-2。实际对生存数据作分布拟合时，可用上述模型分别进行拟合，根据拟合优度检验的结果选择适当的模型。有时，对于一批生存数据，事先不知道生存时间分布的总体趋势，也不好判断应该用什么样的模型最合适，许多研究者一般直接采用非参数方法或半参数法。但是，如果一批数据确实符合某特定的参数模型，由于非参数方法的精度一般低于参数方法，因此，按照非参数方法进行的分析就不能有效地利用和阐述样本数据所包含的信息，同时它对样本量的要求也高于参数方法。

表 26-2 常见生存时间分布的概率密度函数 $f(t)$ 、生存函数 $S(t)$ 和风险函数 $h(t)$

| 分 布 | $f(t)$ | $S(t)$ | $h(t)$ |
|----------------|---|---|---|
| 指数分布 | $\lambda \exp(-\lambda t)$ | $\exp(-\lambda t)$ | λ |
| Weibull 分布 | $\lambda \gamma (\lambda t)^{\gamma-1} \exp[-(\lambda t)^\gamma]$ | $\exp[-(\lambda t)^\gamma]$ | $\lambda \gamma (\lambda t)^{\gamma-1}$ |
| Gamma 分布 | $\frac{\lambda^\gamma t^{\gamma-1} \exp(-\lambda t)}{\Gamma(\gamma)}$ | $1 - I(\lambda t, \gamma)$ | $\frac{f(t)}{S(t)}$ |
| 对数正态分布 | $\frac{\exp\left[-\frac{1}{2}\left(\frac{\ln t - \mu}{\sigma}\right)^2\right]}{\sigma t \sqrt{2\pi}}$ | $1 - \Phi\left[\frac{\ln t - \mu}{\sigma}\right]$ | $\frac{f(t)}{S(t)}$ |
| 对数 logistic 分布 | $\frac{(t\gamma)^{\gamma-1} \lambda}{[1 + (\lambda t)^\gamma]^2}$ | $\frac{1}{1 + (\lambda t)^\gamma}$ | $\frac{(t\gamma)^{\gamma-1} \lambda}{1 + (\lambda t)^\gamma}$ |
| 广义 Gamma 分布 | $\frac{\gamma \lambda^k t^{\gamma k-1} \exp(-\lambda t^\gamma)}{\Gamma(k)}$ | $1 - I[\lambda t^\gamma, k]$ | $\frac{f(t)}{S(t)}$ |

26.3 生存资料参数模型回归分析的 SAS 实现

26.3.1 SAS 程序

考虑到例 26-1 中收缩压和舒张压两个变量有一定的相关性，数据分析时取平均血压 (MBP)，即令 $MBP = SBP * (1/3) + DBP * (2/3)$ 。程序名为 SASTJFX26_1.SAS。

```
DATA SASTJFX26_1;
INPUT Id status t Age1 BMI Age0
smk SBP DBP ECG CHD;
MBP = SBP* (1/3) + DBP* (2/3);
CARDS;
  1 0 12.4 44 34.2 41 0 132 96 0 0
  2 0 12.4 49 32.6 48 2 130 72 0 0
  .....
149 0 10.5 49 30.8 47 1 146 86 0 0
;
run;
```

```
PROC LIFEREG;
CLASS CHD smk ECG;
MODEL t* status(0) =Age1 BMI Age0
MBP CHD smk ECG/ DIST = GAMMA;
RUN;
```

```

ods html;
PROC LIFEREG;
CLASS CHD smk ECG;
MODEL t* status(0) = Age1 BMI Age0
MBP CHD smk ECG/DIST = EXPONENTIAL;
PROC LIFEREG;
CLASS CHD smk ECG;
MODEL t* status(0) = Age1 BMI Age0
MBP CHD smk ECG/ DIST = WEIBULL;

PROC LIFEREG;
CLASS ECG;
MODEL t* status(0) = Age1 MBP
ECG/ DIST = GAMMA;
OUTPUT OUT = a P = median STD = s;
RUN;
PROC PRINT DATA = a;
VAR t status_prob median s;
RUN;
ods html close;

```

【程序说明】 程序第一步建立数据集。为节省篇幅，在 CARDS 语句后的数据仅列出三个观测，其他以省略号表示。第一、二、三个过程步分别对此资料拟合指数模型、Weibull 模型和广义 Gamma 模型回归分析，class 语句中列出离散型变量，model 语句中列出连续型和离散型两类变量，model 语句左边为生存时间和生存结局变量（括号内为截尾值），右边为协变量。根据三个模型拟合效果的评价，选用合适的模型，并在第四个过程步中仅保留有统计学意义的变量，重新进行分析。所得结果输出到数据集 a 中，“P = median”要求将估计的中位生存时间存放到数据集 a 中，“STD = s”为存放分位数标准误的估计结果或 x' b 的标准误估计结果的变量命名。最后打印数据集 a 中的一些变量，包括患者的生存时间、生存结局、估计的分位数、估计的中位生存时间等。

26.3.2 主要分析结果及解释

以下是 SASTJFX26_1.SAS 输出的结果及其解释。

指数模型输出结果：

| Fit Statistics | | | | | | | |
|--|-----|----------|----------------|-----------------------|------------|------------|---|
| - 2 Log Likelihood | | | | 113.645 | | | |
| AIC(smaller is better) | | | | 133.645 | | | |
| AICC(smaller is better) | | | | 135.251 | | | |
| BIC(smaller is better) | | | | 163.617 | | | |
| Analysis of Maximum Likelihood Parameter Estimates | | | | | | | |
| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | Chi-Square | Pr > ChiSq | |
| Intercept | 1 | 10.1918 | 2.3857 | 5.5160 14.8677 | 18.25 | <.0001 | |
| Age1 | 1 | -0.0926 | 0.0349 | -0.1610 -0.0241 | 7.02 | 0.0080 | |
| BMI | 1 | 0.0315 | 0.0331 | -0.0334 0.0964 | 0.91 | 0.3410 | |
| Age0 | 1 | 0.0232 | 0.0309 | -0.0373 0.0837 | 0.57 | 0.4519 | |
| MBP | 1 | -0.0395 | 0.0181 | -0.0749 -0.0040 | 4.77 | 0.0290 | |
| CHD | 0 1 | -0.9838 | 1.0803 | -3.1011 1.1335 | 0.83 | 0.3624 | |
| CHD | 1 0 | 0.0000 | . | . | . | . | . |
| smk | 0 1 | 0.0669 | 0.6392 | -1.1860 1.3198 | 0.01 | 0.9167 | |
| smk | 1 1 | 0.0417 | 0.6553 | -1.2426 1.3259 | 0.00 | 0.9493 | |
| smk | 2 0 | 0.0000 | . | . | . | . | . |
| ECG | 0 1 | 2.1067 | 1.0970 | -0.0433 4.2567 | 3.69 | 0.0548 | |
| ECG | 1 1 | 0.8134 | 0.6410 | -0.4429 2.0697 | 1.61 | 0.2044 | |
| ECG | 2 0 | 0.0000 | . | . | . | . | . |

| | | | | | |
|---------------|---|--------|--------|--------|--------|
| Scale | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 |
| Weibull Shape | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 |

Weibull 模型输出结果：

| Fit Statistics | | | | | | | |
|--|-----|--------------------------|----------------|-----------------------|------------|------------|---|
| | | - 2 Log Likelihood | | 85.294 | | | |
| | | AIC(smaller is better) | | 107.294 | | | |
| | | AICC(smaller is better) | | 109.235 | | | |
| | | BIC(smaller is better) | | 140.263 | | | |
| Analysis of Maximum Likelihood Parameter Estimates | | | | | | | |
| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | Chi-Square | Pr > ChiSq | |
| Intercept | 1 | 5.8033 | 1.0679 | 3.7103 7.8964 | 29.53 | <.0001 | |
| Agel | 1 | -0.0554 | 0.0151 | -0.0850 -0.0258 | 13.45 | 0.0002 | |
| BMI | 1 | 0.0075 | 0.0119 | -0.0158 0.0308 | 0.40 | 0.5289 | |
| Age0 | 1 | 0.0225 | 0.0128 | -0.0027 0.0476 | 3.07 | 0.0796 | |
| MBP | 1 | -0.0159 | 0.0075 | -0.0306 -0.0013 | 4.54 | 0.0330 | |
| CHD | 0 1 | -0.5804 | 0.4016 | -1.3674 0.2067 | 2.09 | 0.1484 | |
| CHD | 1 0 | 0.0000 | . | . | . | . | . |
| smk | 0 1 | 0.1661 | 0.2396 | -0.3035 0.6357 | 0.48 | 0.4881 | |
| smk | 1 1 | 0.0499 | 0.2481 | -0.4363 0.5361 | 0.04 | 0.8405 | |
| smk | 2 0 | 0.0000 | . | . | . | . | . |
| ECG | 0 1 | 1.2240 | 0.4269 | 0.3874 2.0607 | 8.22 | 0.0041 | |
| ECG | 1 1 | 0.2908 | 0.2543 | -0.2076 0.7892 | 1.31 | 0.2528 | |
| ECG | 2 0 | 0.0000 | . | . | . | . | . |
| Scale | 1 | 0.3490 | 0.0555 | 0.2555 0.4767 | | | |
| Weibull Shape | 1 | 2.8653 | 0.4559 | 2.0977 3.9139 | | | |

广义 Gamma 模型输出结果：

| Fit Statistics | | | | | | | |
|----------------|--|--------------------------|--|---------|--|--|--|
| | | - 2 Log Likelihood | | 73.255 | | | |
| | | AIC(smaller is better) | | 97.255 | | | |
| | | AICC(smaller is better) | | 99.566 | | | |
| | | BIC(smaller is better) | | 133.221 | | | |

WARNING: The relative gradient convergence criterion of 0.0399615291 is greater than the limit of 0.0001. The convergence is questionable.

| Analysis of Maximum Likelihood Parameter Estimates | | | | | | | |
|--|-----|----------|----------------|-----------------------|------------|------------|---|
| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | Chi-Square | Pr > ChiSq | |
| Intercept | 1 | 4.5667 | 0.8481 | 2.9044 6.2290 | 28.99 | <.0001 | |
| Agel | 1 | -0.0491 | 0.0114 | -0.0715 -0.0267 | 18.44 | <.0001 | |
| BMI | 1 | 0.0265 | 0.0115 | 0.0039 0.0492 | 5.29 | 0.0214 | |
| Age0 | 1 | 0.0086 | 0.0091 | -0.0092 0.0264 | 0.89 | 0.3453 | |
| MBP | 1 | -0.0126 | 0.0068 | -0.0259 0.0008 | 3.41 | 0.0649 | |
| CHD | 0 1 | -0.4816 | 0.2681 | -1.0070 0.0438 | 3.23 | 0.0724 | |
| CHD | 1 0 | 0.0000 | . | . | . | . | . |
| smk | 0 1 | 0.3033 | 0.1698 | -0.0295 0.6361 | 3.19 | 0.0741 | |

| | | | | | | | | |
|-------|---|---|---------|--------|---------|--------|------|--------|
| smk | 1 | 1 | 0.3415 | 0.2037 | -0.0578 | 0.7408 | 2.81 | 0.0937 |
| smk | 2 | 0 | 0.0000 | . | . | . | . | . |
| ECG | 0 | 1 | 0.8423 | 0.2843 | 0.2851 | 1.3995 | 8.78 | 0.0030 |
| ECG | 1 | 1 | -0.0306 | 0.3368 | -0.6907 | 0.6295 | 0.01 | 0.9276 |
| ECG | 2 | 0 | 0.0000 | . | . | . | . | . |
| Scale | | 1 | 0.3843 | 0.1632 | 0.1671 | 0.8834 | | |
| Shape | | 1 | -3.1504 | 1.9357 | -6.9443 | 0.6434 | | |

在 PROC LIFEREG 中没有直接拟合标准 Gamma 模型的方法，但 PROC LIFEREG 可以将尺度参数和形状参数设定为特定值。若拟合标准 Gamma 模型，可试用许多不同的值(比如用直线搜索法)，直到找到一个能使对数似然值达到最大的共同的尺度参数和形状参数。本例不再尝试，感兴趣的读者可以参阅有关书籍尝试进行。

现比对三个模型的拟合效果，可采用似然比检验，似然比统计量的公式为：

$$\chi^2_v = -2\log L_q - (-2\log L_{q+v})$$

式中 χ^2_v 服从自由度为 v 的 χ^2 分布， $-2\log L_q$ 和 $-2\log L_{q+v}$ 分别为含 q 和 $q+v$ 个参数的模型的对数似然函数值。因 Weibull 分布包含 2 个参数，指数分布包含 1 个参数，广义 Gamma 分布包含 3 个参数，标准 Gamma 分布包含 2 个参数。所以，各种分布拟合效果有无差异的假设检验结果可汇总成如表 26-3 所示。

表 26-3 糖尿病资料似然比拟合优度检验

| 比较模型 | | | 自由度 | χ^2 统计量 | P 值 |
|------------|----|-------------|-----|--------------|------------|
| 指数模型 | VS | Weibull 模型 | 1 | 28.351 | $P < 0.05$ |
| 指数模型 | VS | 广义 Gamma 模型 | 2 | 40.390 | $P < 0.05$ |
| Weibull 模型 | VS | 广义 Gamma 模型 | 1 | 12.039 | $P < 0.05$ |

可见，各分布的拟合效果之间均有统计学差异。故这里应选用 $-2\log L$ 值最小的分布，即广义 Gamma 分布($-2\log L$ 值为 73.255)。查看广义 Gamma 模型的输出结果，只有 Age1、MBP 和 ECG 三个变量有统计学差异。故仅保留这三个变量，重新采用广义 Gamma 模型来进行分析。所得结果如下：

| Analysis of Maximum Likelihood Parameter Estimates | | | | | | | |
|--|---|----|----------|----------------|-----------------------|------------|------------|
| Parameter | | DF | Estimate | Standard Error | 95% Confidence Limits | Chi-Square | Pr > ChiSq |
| Intercept | | 1 | 5.8937 | 0.8060 | 4.3140 7.4734 | 53.47 | <.0001 |
| Age1 | | 1 | −0.0341 | 0.0065 | −0.0468 −0.0214 | 27.73 | <.0001 |
| MBP | | 1 | −0.0177 | 0.0063 | −0.0301 −0.0053 | 7.88 | 0.0050 |
| ECG | 0 | 1 | 0.7571 | 0.2518 | 0.2635 1.2506 | 9.04 | 0.0026 |
| ECG | 1 | 1 | 0.3760 | 0.2821 | −0.1769 0.9290 | 1.78 | 0.1826 |
| ECG | 2 | 0 | 0.0000 | . | . | . | . |
| Scale | | 1 | 0.5865 | 0.1094 | 0.4069 0.8453 | | |
| Shape | | 1 | −0.5880 | 0.9093 | −2.3701 1.1941 | | |

广义 Gamma 回归模型结果表明，随访开始时年龄和平均血压的回归系数均为负值(-0.0341 和 -0.0177)，说明这两个变量取值越大，生存时间越短，死亡风险越大；心电图读数的 0 水平与 2 水平比较时，ECG0 对应的 P 值为 0.0026，且其回归系数为正值(0.7571)，说明心电图正常者生存时间长于异常者。

以下是基于广义 Gamma 模型预测各位患者中位生存时间的结果(为节省篇幅,仅给出前5位患者的相关结果)。

| Obs | t | status | _PROB_ | median | s |
|-----|------|--------|--------|---------|---------|
| 1 | 12.4 | 0 | 0.5 | 28.7227 | 6.4425 |
| 2 | 12.4 | 0 | 0.5 | 32.5422 | 7.1914 |
| 3 | 9.6 | 0 | 0.5 | 43.7178 | 11.8530 |
| 4 | 7.2 | 0 | 0.5 | 22.9713 | 6.1848 |
| 5 | 14.1 | 0 | 0.5 | 33.8405 | 7.5489 |

前5位患者的输出结果表明,第1号患者的随访开始时年龄(Age1)为44岁,收缩压(SBP)为132mmHg,舒张压(DBP)为96 mmHg,心电图读数正常,预测此类患者的中位生存时间为28.7年。第2号患者的随访开始时年龄(Age1)为49岁,收缩压(SBP)为130mmHg,舒张压(DBP)为72mmHg,心电图读数正常,预测此类患者的中位生存时间为32.5年。其他依此类推。

26.4 LIFEREG 过程简介

PROCLIFEREG 过程对生存数据拟合参数模型,其最大特点在于可以处理右截尾、左截尾或区间截尾数据,同时含有丰富的生存分布形式,特别是其中的广义 Gamma 分布可以进行许多其他概率分布的似然比拟合优度检验。

CLASS 语句用于说明分类变量。

MODEL 语句指出哪些变量用于该模型的回归部分以及模型的误差项或随机项的分布是什么。MODEL 语句可用的选项有:

(1) DISTRIBUTION|DIST|D = distribution - type(分布的类型),说明生存时间的分布类型。exponential、weibull、gamma、normal、lnormal、logistic、llogistic 指定指数分布、Weibull 分布、Gamma 分布、正态分布、对数正态分布、logistic 分布和对数 logistic 分布。

(2) NOLOG 要求不对反应变量进行对数变换,缺省时, LIFEREG 过程对反应变量进行对数变换。

(3) SCALE = value(值),要求尺度参数以这个值作为初始值。

(4) NOSCALE 要求尺度参数固定。

(5) SHAPE1 = value(值),要求形状参数用规定的 value 值为初始值。

(6) NOSHAPE1 要求第一个形状参数 SHAPE1 保持固定。

OUTPUT 语句创建一个新 SAS 数据集,它包含模型拟合之后计算的统计量。

OUTPUT < OUT = SAS-data-set > keyword = name;

OUT = SAS-data-set(SAS 数据集),命名输出数据集。keyword = name(关键词 = 名字),规定在 OUTPUT 数据集中包含的统计量(如下),并给出包含这些统计量的新名字。

(1) CONTROL 在输入数据集中命名用于控制分位数估计的变量。

(2) PREDICTED|P,命名存放分位数估计结果的变量。缺省时计算第50百分位数即中位生存时间。

(3) QUANTILES|Q,给出所要求计算的分位数列表。

(4) STD_ERR|STD,命名存放分位数标准差估计结果的变量。

(5)XBETA 命名存放 $x'b$ 计算结果的变量。

LIFEREG 过程也可以得到原始数据集外其他协变量值所对应的预测值。模型拟合前, 将这些协变量值附加在原数据集后, 生存时间设置为缺失值。这样这些观测不用于模型拟合, 但可生成它们的预测值。如果只需要几个观察值(真实值或假想值)的预测, 则生成一个变量(如 USE), 若需要预测, 则该变量等于 1; 否则, 该变量等于 0。OUTPUT 语句中包括 `CONTROL = USE`。

26.5 本章小结

(1)本章主要介绍了生存资料的参数模型回归分析, 并举例说明如何借助 SAS 软件的 LIFEREG 过程完成此分析。另外, 还简要介绍了 LIFEREG 过程的常见选项及其含义, 方便读者根据需要加以选用。

(2)当资料服从特定分布(如指数分布、Weibull 分布或 Gamma 分布等)时应选用相应的参数回归模型, 似然比检验可用于参数回归模型的拟合优度检验。建立相应模型后, 可估计中位生存时间以评价预后。

(余红梅 高辉)

第 27 章 时间序列分析

本章主要介绍如何分析随时间变化的数据之间的内在关系,其研究领域被称为时间序列分析。无论是从概念还是从应用上看,时间序列分析与通常的回归分析都有很大的区别,但又很像回归分析。对于它需要哪些基础、它能解决哪些医学研究领域中的实际问题,将在本章中给出详细解答。

27.1 时间序列分析简介

在医学研究中,按某种(相等或不相等)的时间间隔对客观事物进行动态观察,由于随机因素的影响,各次观察的指标 $x_1, x_2, x_3, \dots, x_i, \dots$ 都是随机变量,这种按时间顺序排列的随机变量的一组实测值称为时间序列。例如生命体征监护获得的随时间变化的一系列读数,某地区某种疾病发病率或死亡率等指标的定期监测数据均构成时间序列。由于随机因素的作用,各时刻的观测可视作一随机变量,当 $t \in (a, b)$, 则变量集合 $\{x_t\}$ 称为随机过程,实际工作中的实测值序列称为随机过程的一次实现。参数 t 可以是时间,也可以是其他有序变量,如空间位置、温度水平等,甚至可以是向量。

时间序列中每一时期的数值,都是由许多不同的因素共同作用的结果。而这些因素往往交织在一起,这样就增加了分析时间序列的困难。因此,时间序列分析通常对各种可能发生作用的因素进行分类,如长期趋势、季节变动、循环变动和不规则变动。

时间序列分析的目的是利用所拟合的模型对某研究领域中的动态数据的未来状况进行预测。时间序列分析大致包括三方面的内容:(1)选择模型并进行参数估计;(2)模型的适用性检验;(3)预测预报。

时间序列分析模型是统计学学科中发展迅速的一个分支,本章主要介绍时域分析、频域分析中的基本原理与方法;对于更全面的学习,需要参考时间序列理论与方法的相关专著。

27.2 指数平滑法

27.2.1 指数平滑法简介

指数平滑是(Exponential smoothing)由 Brown 等(Brown 和 Meyers 1961; Brown 1962)发展起来的计算模式,它拟合一种使用平滑方案的时间趋势模型。通式是 $S_t = \alpha x_t + (1 - \alpha)S_{t-1}$, 式中 S_t 为第 t 期平滑值($t > 0$), α 为平滑系数(取值范围 $0 < \alpha < 1$), x_t 为第 t 期实际观察值,系数 α 和 $(1 - \alpha)$ 都表示权重。在此方案里面权重大小随着时间的向后推移而呈几何级数下降。因为对于事物未来发展的水平,新近的观测值比早期的观测值的预测价值更大,因而在预测时,新近观测值应比早期观测值具有更大的权重。作为一种预测方法,指数平滑预测效果的好坏取决于对这个序列选择一个怎样的平滑系数 α 。 α 值界于 $0 \sim 1$ 之间。一般来说,平滑系数 α 的取值

大小应当视预测对象的特点及预测周期的长短而定。 α 取值偏低时, 预测结果主要取决于历史情形, 不能及时跟踪数据新的变化趋势; α 取值偏高时, 预测模型具有较高的灵敏度, 能够迅速跟踪新数据的变化, 但对历史数据的信息利用较少。实际应用中, 通常采用多个水平的 α 值进行试算比较, 选择其中的最优值作为平滑系数, 原则是使预期预测误差平方和 (SSE)、平均平方误差 (MSE) 或平均绝对误差 (MAE) 最小。在根据上述原则进行优选后, 还应该对根据预测结果所得到的参数的合理性进行检验。

27.2.2 应用实例

【例 27-1】某药品公司自 1990 年以来生产的某种抗生素的出厂数量时间序列如表 27-1 所示。试用指数平滑法预测 2006—2010 年该药的出厂数量, 并计算模型参数、预测值及其置信区间。

表 27-1 某药品公司 1990—2005 年
某种抗生素的出厂数量

| 年 份 | 数 量 |
|------|-------|
| 1990 | 371.5 |
| 1991 | 267.4 |
| ... | ... |
| 2005 | 336.4 |

27.2.3 SAS 程序

SAS 程序名为 sastjfx27_1.sas。

```

data;
input number @@ ;
year=intnx('year','1jan1990'd,_n_
-1);
format year year4.;
cards;
371.5 267.4 372.4 368.2 349.4
362.8 420.9 380.4 385.6 335.0
338.5 306.6 323.6 341.3 354.2
336.4
;
run;
ods html;

proc forecast data=lead=5
interval=year method=expo
trend=2 weight=0.2 out=pred
outlimit outest=est outfitstats;
id year;
var number;
run;
proc print data=pred;
run;
proc print data=est;
run;
ods html close;

```

【程序说明】 第一步是数据步, 用来创建待分析的数据集。数据集中有两个变量, 变量 year 为日期变量, 变量 number 表示从 1990 到 2005 年的出厂数量, “year = intnx(' year ', '1jan1990'd, _n_ - 1);”该语句表示从 1990 年 1 月 1 日起, 按年递增日期, 自动变量 _n_ 计算 DATE 步程序已执行了多少步, 因此 _n_ - 1 为从第一个观测日期算起所需增加的周期数。“format year year4.”表示输出日期值中的年份。第二步调用 forecast 过程进行时间序列分析。lead = 5 表示向前预测 5 个周期, interval = year 表示预测观察的 SAS 日期值按年外推得到, method = expo 表示利用指数平滑法进行预测, trend = 2 表示拟合的是线性趋势模型, weight = 0.2 规定平滑系数为 0.2, out = pred 表示将预测结果输出到名为 pred 的数据集中, outlimit 表示计算预测值的 95% 置信区间 (包括上限和下限); outest = est 表示将模型的参数值输出到名为 est 的数据集中, outfitstats 表示将不同的 R^2 类型预测精度统计量输出到 est 数据集中。id year 表示用含有 SAS 日期值的变量 year 识别观测; var number 表示对输入数据集中的 number 进行预测。

27.2.4 主要分析结果及解释

以下是计算所得的预测值及其置信区间。

| Obs | year | _TYPE_ | _LEAD_ | number |
|-----|------|----------|--------|---------|
| 1 | 2006 | FORECAST | 1 | 343.220 |
| 2 | 2006 | L95 | 1 | 238.739 |
| 3 | 2006 | U95 | 1 | 447.701 |
| 4 | 2007 | FORECAST | 2 | 343.516 |
| 5 | 2007 | L95 | 2 | 235.145 |
| 6 | 2007 | U95 | 2 | 451.887 |
| ... | ... | ... | ... | ... |
| 13 | 2010 | FORECAST | 5 | 344.403 |
| 14 | 2010 | L95 | 5 | 221.391 |
| 15 | 2010 | U95 | 5 | 467.416 |

在输出的数据集 pred 中, 5 个预测周期中的每一个都包括了三个观测, 而这三个观测的 ID 变量值 year 都是相同的, 也就是该观测的年份值。每个预测周期的三个观测对应于变量 _TYPE_ 有不同的值, _TYPE_ = FORECAST 的观测中变量 number 的值是该周期的预测值; _TYPE_ = L95 的观测中变量 number 的值是该周期预测值 95% 置信区间的下限; _TYPE_ = U95 的观测中变量 number 的值是该周期预测值 95% 置信区间的上限。

以下是模型参数的计算结果。

| Obs | _TYPE_ | year | number | Obs | _TYPE_ | year | number |
|-----|----------|------|-----------|-----|---------|------|-----------|
| 1 | N | 2005 | 16 | 16 | MAE | 2005 | 33.38561 |
| 2 | NRESID | 2005 | 16 | 17 | ME | 2005 | -14.08767 |
| 3 | DF | 2005 | 14 | 18 | MAXE | 2005 | 44.240131 |
| 4 | WEIGHT | 2005 | 0.2 | 19 | MINE | 2005 | -87.9319 |
| 5 | S1 | 2005 | 341.74109 | 20 | MAXPE | 2005 | 11.431135 |
| 6 | S2 | 2005 | 340.5579 | 21 | MINPE | 2005 | -32.88403 |
| 7 | SIGMA | 2005 | 44.2694 | 22 | RSQUARE | 2005 | -0.466901 |
| 8 | CONSTANT | 2005 | 342.92428 | 23 | ADJRSQ | 2005 | -0.57168 |
| 9 | LINEAR | 2005 | 0.2957969 | 24 | RW_RSQ | 2005 | 0.1974124 |
| 10 | SST | 2005 | 18703.997 | 25 | ARSQ | 2005 | -0.886016 |
| 11 | SSE | 2005 | 27436.917 | 26 | APC | 2005 | 2204.7522 |
| 12 | MSE | 2005 | 1959.7798 | 27 | AIC | 2005 | 123.1529 |
| 13 | RMSE | 2005 | 44.2694 | 28 | SBC | 2005 | 124.69807 |
| 14 | MAPE | 2005 | 10.021207 | 29 | CORR | 2005 | 0.1539706 |
| 15 | MPE | 2005 | -4.978152 | | | | |

在 est 数据集中, 变量 year 包含用于拟合预测模型的数据集 exp1 中最后的观测的 ID 值 (2005)。变量 number 提供了 _TYPE_ 列各统计量的值。_TYPE_ = N、NRESID 和 DF 分别表示来自数据集的观测个数、残差个数和自由度; _TYPE_ = WEIGHT 表示平滑系数; _TYPE_ = S1 表示一次指数平滑的最后平滑值; _TYPE_ = S2 表示二次指数平滑的最后平滑值; _TYPE_ = SIGMA 表示预测误差方差的估计值; _TYPE_ = CONSTANT 和 LINEAR 分别给出了时间趋势模型的截距和回归系数; _TYPE_ = SST 表示对均值修正的总平方和; _TYPE_ = SSE、MSE 和 MAE 分别表示预测误差平方和、均方误差和平均绝对误差; _TYPE_ = RMSE 表示均方根误差; _TYPE_ = MAPE 表示平均绝对百分数误差; _TYPE_ = MPE 表示平均百分数误差; _TYPE_ = ME 表示平均误差; TYPE = MAXE、MINE 表示最大误差 (最大残差)、最小误差 (最小残差);

`_TYPE_ = MAXPE`、`MINPE` 表示最大百分数误差、最小百分数误差；`_TYPE_ = RSQUARE` 表示 R^2 统计量；`_TYPE_ = ADJRSQ`、`RW_RSQ`、`ARSQ` 分别表示修正的 R^2 统计量、随机游动 R^2 统计量、Amemiya 修正的 R^2 统计量；`_TYPE_ = APC` 表示 Amemiya 预报准则；`_TYPE_ = AIC` 表示 Akaike 信息准则；`_TYPE_ = SBC` 表示 Schwarz 贝叶斯准则；`_TYPE_ = CORR` 表示实测值与向前一步预报值之间的相关系数。

27.3 ARIMA 模型

27.3.1 ARIMA 模型简介

ARIMA 方法最先由 Box 和 Jenkins 大量使用，所以 ARIMA 模型通常也称为 Box-Jenkins 模型。Box-Tiao(1975) 讨论了 ARIMA 过程所使用的转移函数模型。当 ARIMA 模型包括其他时间序列作为输入变量时，该模型有时被称为 ARIMAX 模型。Pankratz(1991) 称 ARIMAX 模型为“动态回归”。

SAS 软件中 ARIMA 过程提供了一个综合工具包来进行一元时间序列的模型识别、参数估计以及预测。利用 ARIMA 模型或 ARMA 模型，ARIMA 过程分析并预测等间距一元时间序列数据、转移函数数据、干预数据。ARIMA 模型通过其自身的过去值、过去误差、其他时间序列的当前值和过去值的线性组合来预测响应时间序列。ARIMA 过程支持如下模型：季节、子集和因子 ARIMA 模型；干预或中断时间序列模型；带 ARIMA 误差的多元回归分析；任意复杂程度的有理转移函数模型。

27.3.2 应用实例

【例 27-2】 为了对某区级医院体检中心 1970—2005 年的收入序列(连续性变量)进行预测(数据详见表 27-2)，选用 ARIMA 过程进行。通过 ARIMA 模型预测的三个主要阶段，通过识别(自相关、逆自相关及偏自相关函数，并最终根据 AIC 最小的原则选择合适的模型)、建模和预测，完成对该医院收入序列的预测结果。

表 27-2 某区级医院体检中心 1970—2005 年的收入序列

| 年度 | 收入 | 年度 | 收入 | 年度 | 收入 |
|------|---------|------|---------|------|---------|
| 1970 | 2575800 | 1982 | 2953430 | 1994 | 3329890 |
| 1971 | 2606680 | 1983 | 2986450 | 1995 | 3361130 |
| ... | ... | ... | ... | ... | ... |
| 1981 | 2921900 | 1993 | 3297650 | 2005 | 3674310 |

27.3.3 SAS 程序

SAS 程序名为 `sastjfx27_2.sas`。

```
data;
input a1 @@ ;
year=intnx('year','1jan1970'd,
           _n_-1);
format year year4.;
cards;
2575800 2606680 2639000 2671000
2702380 2733890 2765840 2796710
2829000 2859800
.....
```

```
plot a1* year;
run;
proc arima data=expl;
identify var=a1 nlag=16;
identify var=a1 (1) nlag=16;
/* 一阶差分后序列的识别*/
estimate p=2;
forecast lead=6 interval=year id
=year out=out;
run;
```

```

3674310
run;
ods html;
proc gplot;
symbol i = spline v = star h = 2 c =
black;
ods html close;

```

【程序说明】 第一步是数据步，用来创建名为 exp1 的 SAS 数据集。第二步调用 gplot 过程，用来绘制数据与时间的关系图，初步识别序列的样式。第三步调用 arima 过程，以识别模型，并根据自相关、偏自相关函数的截尾和拖尾性质，选择合适的估计模型。identify 语句指定被建模的时间序列，“var = a1”指定需分析的时间序列的变量为 a1，“var = a1(1)”表示对变量 a1 进行一阶差分(括号内的数字表示拆分的间隔数)，“nlag = 16”指明计算自相关系数和互相关系数过程中所需考虑的时间间隔个数。estimate 语句对前面的 identify 语句中指定的响应序列拟合 ARMA 模型或转移函数模型，并产生其参数估计，“p = 2”指定模型的自回归部分。forecast 语句使用前面 estimate 语句生成的参数估计来产生时间序列的预测值。

27.3.4 主要分析结果及解释

| Autocorrelations | | | | | | | | | | | | | | | | | | | | | | | | |
|------------------|------------|-------------|----|---|---|---|---|---|---|---|---|---|-----|---|-------|---|---|---|---|---|---|---|---|-----------|
| Lag | Covariance | Correlation | -1 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 1 | Std Error |
| 0 | 1.06171E11 | 1.00000 | | | | | | | | | | | | | ***** | | | | | | | | | 0 |
| 1 | 9.73114E10 | 0.91655 | | | | | | | | | | | | | ***** | | | | | | | | | 0.166667 |
| 2 | 8.84519E10 | 0.83311 | | | | | | | | | | | | | ***** | | | | | | | | | 0.272852 |
| 3 | 7.96815E10 | 0.75050 | | | | | | | | | | | | | ***** | | | | | | | | | 0.336166 |
| 4 | 7.09958E10 | 0.66869 | | | | | | | | | | | | | ***** | | | | | | | | | 0.379868 |
| 5 | 6.24009E10 | 0.58774 | | | | | | | | | | | | | ***** | | | | | | | | | 0.411268 |
| 6 | 5.39802E10 | 0.50843 | | | | | | | | | | | | | ***** | | | | | | | | | 0.433972 |
| 7 | 4.57058E10 | 0.43049 | | | | | | | | | | | | | ***** | | | | | | | | | 0.450215 |
| 8 | 3.76264E10 | 0.35439 | | | | | | | | | | | | | ***** | | | | | | | | | 0.461507 |
| 9 | 2.97834E10 | 0.28052 | | | | | | | | | | | | | ***** | | | | | | | | | 0.469006 |
| 10 | 2.21801E10 | 0.20891 | | | | | | | | | | | | | **** | | | | | | | | | 0.473644 |
| 11 | 1.48606E10 | 0.13997 | | | | | | | | | | | | | *** | | | | | | | | | 0.476196 |
| 12 | 7814011734 | 0.07360 | | | | | | | | | | | | | * | | | | | | | | | 0.477338 |
| 13 | 1112136804 | 0.01047 | | | | | | | | | | | | | | | | | | | | | | 0.477653 |
| 14 | -5.21792E9 | -.04915 | | | | | | | | | | | * | | | | | | | | | | | 0.477659 |
| 15 | -1.1181E10 | -.10531 | | | | | | | | | | | ** | | | | | | | | | | | 0.477800 |
| 16 | -1.6748E10 | -.15774 | | | | | | | | | | | *** | | | | | | | | | | | 0.478444 |

"," marks two standard errors

| Inverse Autocorrelations | | | | | | | | | | | | | | | | | | | | | | | | |
|--------------------------|-------------|----|---|---|---|---|---|---|---|---|---|---|-----|---|---|---|---|---|---|---|---|---|--|--|
| Lag | Correlation | -1 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 1 | | |
| 1 | -0.50125 | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | 0.00282 | | | | | | | | | | | | | | | | | | | | | | | |
| ... | ... | | | | | | | | | | | | ... | | | | | | | | | | | |
| 16 | 0.02235 | | | | | | | | | | | | | | | | | | | | | | | |

| Partial Autocorrelations | | | | | | | | | | | | | | | | | | | | | | | | |
|--------------------------|-------------|----|---|---|---|---|---|---|---|---|-----|---|---|-------|---|---|---|---|---|---|---|---|--|--|
| Lag | Correlation | -1 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 1 | | |
| 1 | 0.91655 | | | | | | | | | | | | | ***** | | | | | | | | | | |
| 2 | -0.04354 | | | | | | | | | | * | | | | | | | | | | | | | |
| ... | ... | | | | | | | | | | ... | | | | | | | | | | | | | |
| 16 | -0.04178 | | | | | | | | | | * | | | | | | | | | | | | | |

以上分别是绘制序列随时间变化的时序图, 序列的自相关、逆自相关和偏自相关函数分析结果, 可帮助选择合适的 ARMA 模型。由图 27-1 可见, 时序图有较为明显的线性趋势, 且自相关函数分析结果呈现缓慢的指数函数递减, 所以该序列可以认为是一个非平稳序列, 需要进行一阶差分。逆自相关函数与偏自相关函数分析结果表现为明显的一阶截尾性。

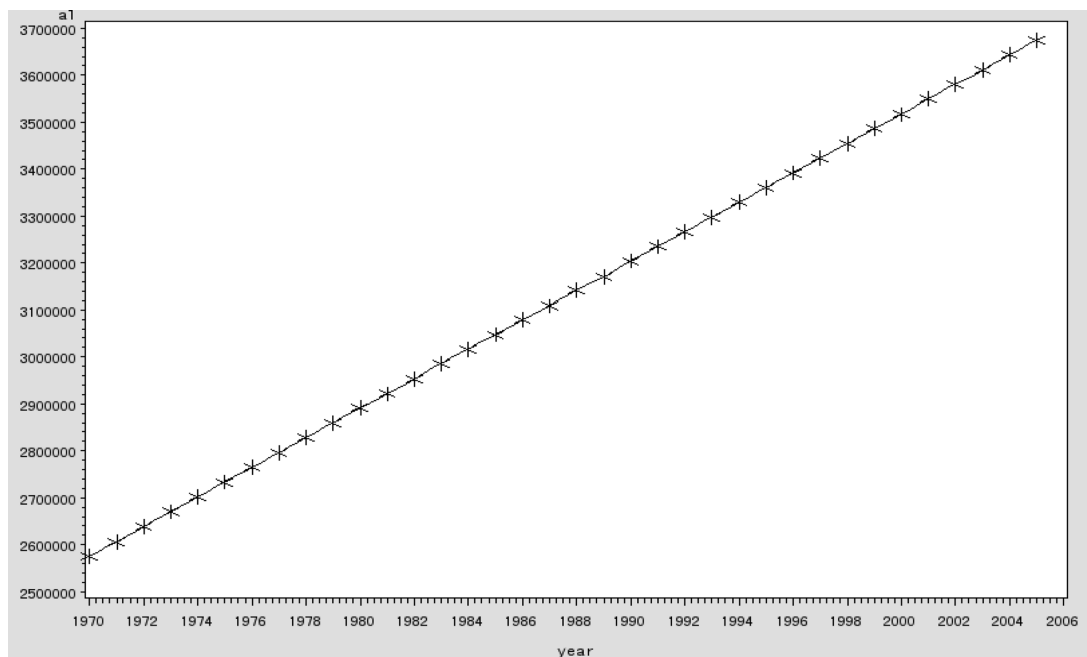


图 27-1 某院体检中心 1970—2005 年收入的时间序列图

以下分别是一阶差分后序列的自相关、逆自相关与偏自相关函数的结果。

| Autocorrelations | | | | | | | | | | | | | | | | |
|------------------|------------|-------------|----|---|---|---|---|---|---|---|---|---|---|-------|---|---|
| Lag | Covariance | Correlation | -1 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 | 1 | 2 | 3 |
| 0 | 1431413 | 1.00000 | | | | | | | | | | | | ***** | | |
| 1 | -974627 | -.68088 | | | | | | | | | | | | | | |
| 2 | 341531 | 0.23860 | | | | | | | | | | | | | | |
| 3 | 17727.200 | 0.01238 | | | | | | | | | | | | | | |
| 4 | -137826 | -.09629 | | | | | | | | | | | | | | |
| 5 | 174019 | 0.12157 | | | | | | | | | | | | | | |
| 6 | -110339 | -.07708 | | | | | | | | | | | | | | |
| 7 | 30816.800 | 0.02153 | | | | | | | | | | | | | | |
| 8 | 123449 | 0.08624 | | | | | | | | | | | | | | |
| 9 | -259474 | -.18127 | | | | | | | | | | | | | | |
| 10 | 102259 | 0.07144 | | | | | | | | | | | | | | |
| 11 | 194980 | 0.13622 | | | | | | | | | | | | | | |
| 12 | -292496 | -.20434 | | | | | | | | | | | | | | |
| 13 | 152059 | 0.10623 | | | | | | | | | | | | | | |
| 14 | -14392.686 | -.01005 | | | | | | | | | | | | | | |
| 15 | -184414 | -.12883 | | | | | | | | | | | | | | |
| 16 | 356385 | 0.24897 | | | | | | | | | | | | | | |

“.” marks two standard errors

| | | Inverse Autocorrelations | | | | | | | | | | | | | | | | | | | | | | | | | | |
|-----|-------------|--------------------------|---|---|---|---|---|---|---|---|------|---|-------|---|---|---|---|---|---|---|---|---|--|--|--|--|--|--|
| Lag | Correlation | -1 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 1 | | | | | | |
| 1 | 0.67122 | | | | | | | | | . | | | ***** | | | | | | | | | | | | | | | |
| 2 | 0.21360 | | | | | | | | | . | | | **** | . | | | | | | | | | | | | | | |
| 3 | -0.03766 | | | | | | | | | . | * | | | | . | | | | | | | | | | | | | |
| 4 | -0.12641 | | | | | | | | | . | *** | | | | . | | | | | | | | | | | | | |
| 5 | -0.17084 | | | | | | | | | . | *** | | | | . | | | | | | | | | | | | | |
| 6 | -0.18651 | | | | | | | | | . | **** | | | | . | | | | | | | | | | | | | |
| 7 | -0.11723 | | | | | | | | | . | ** | | | | . | | | | | | | | | | | | | |
| 8 | 0.04045 | | | | | | | | | . | | | * | . | | | | | | | | | | | | | | |
| 9 | 0.11778 | | | | | | | | | . | | | ** | . | | | | | | | | | | | | | | |
| 10 | 0.05703 | | | | | | | | | . | | | * | . | | | | | | | | | | | | | | |
| 11 | 0.05876 | | | | | | | | | . | | | * | . | | | | | | | | | | | | | | |
| 12 | 0.17063 | | | | | | | | | . | | | *** | . | | | | | | | | | | | | | | |
| 13 | 0.21729 | | | | | | | | | . | | | **** | . | | | | | | | | | | | | | | |
| 14 | 0.15602 | | | | | | | | | . | | | *** | . | | | | | | | | | | | | | | |
| 15 | 0.04962 | | | | | | | | | . | | | * | . | | | | | | | | | | | | | | |
| 16 | -0.01920 | | | | | | | | | . | | | | . | | | | | | | | | | | | | | |

| | | Partial Autocorrelations | | | | | | | | | | | | | | | | | | | | | | | | | | |
|-----|-------------|--------------------------|---|---|---|---|---|---|---|---|-------|---|-------|---|---|---|---|---|---|---|---|---|--|--|--|--|--|--|
| Lag | Correlation | -1 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 1 | | | | | | |
| 1 | -0.68088 | | | | | | | | | | | | ***** | | | . | | | | | | | | | | | | |
| 2 | -0.41948 | | | | | | | | | | | | ***** | | | . | | | | | | | | | | | | |
| 3 | -0.09644 | | | | | | | | | . | ** | | | | . | | | | | | | | | | | | | |
| 4 | -0.06201 | | | | | | | | | . | * | | | | . | | | | | | | | | | | | | |
| 5 | 0.06068 | | | | | | | | | . | | | * | . | | | | | | | | | | | | | | |
| 6 | 0.08538 | | | | | | | | | . | | | ** | . | | | | | | | | | | | | | | |
| 7 | 0.04592 | | | | | | | | | . | | | * | . | | | | | | | | | | | | | | |
| 8 | 0.19320 | | | | | | | | | . | | | **** | . | | | | | | | | | | | | | | |
| 9 | -0.01835 | | | | | | | | | . | | | | . | | | | | | | | | | | | | | |
| 10 | -0.28281 | | | | | | | | | . | ***** | | | | . | | | | | | | | | | | | | |
| 11 | 0.02523 | | | | | | | | | . | | | * | . | | | | | | | | | | | | | | |
| 12 | 0.05610 | | | | | | | | | . | | | * | . | | | | | | | | | | | | | | |
| 13 | -0.08408 | | | | | | | | | . | ** | | | | . | | | | | | | | | | | | | |
| 14 | -0.03781 | | | | | | | | | . | * | | | | . | | | | | | | | | | | | | |
| 15 | -0.23734 | | | | | | | | | . | ***** | | | | . | | | | | | | | | | | | | |
| 16 | 0.05123 | | | | | | | | | . | | | * | . | | | | | | | | | | | | | | |

对差分后的序列求自相关函数和偏自相关函数,可见,样本自相关函数(SACF)呈现拖尾,样本偏自相关函数(SPACF)在 lag = 2 处截尾,故拟合 AR(2) 模型为宜。

Conditional Least Squares Estimation

| Parameter | Estimate | Standard Error | t Value | Approx Pr > t | Lag |
|-----------|----------|----------------|---------|----------------|-----|
| MU | 31403.4 | 57.84464 | 542.89 | <.0001 | 0 |
| AR1,1 | -0.99449 | 0.16036 | -6.20 | <.0001 | 1 |
| AR1,2 | -0.43437 | 0.16320 | -2.66 | 0.0121 | 2 |

| | |
|--------------------|----------|
| Constant Estimate | 76274.64 |
| Variance Estimate | 667697.8 |
| Std Error Estimate | 817.1278 |

| | |
|---------------------|---------|
| AIC | 571.595 |
| SBC | 576.261 |
| Number of Residuals | 35 |

* AIC and SBC do not include log determinant.

这是对数据模型进行最小二乘估计的结果。通过对拟合各阶模型的检验结果发现,均有统计学意义;且 AIC 值不大,可认为模型的拟合效果好。

Autocorrelation Check of Residuals

| To Lag | Chi-Square | DF | Pr > ChiSq | Autocorrelations | | | | | |
|--------|------------|----|------------|------------------|--------|--------|-------|--------|--------|
| 6 | 1.72 | 4 | 0.7866 | -0.039 | -0.091 | -0.043 | 0.114 | 0.111 | 0.066 |
| 12 | 8.85 | 10 | 0.5465 | 0.087 | -0.070 | -0.245 | 0.081 | 0.186 | -0.154 |
| 18 | 14.74 | 16 | 0.5440 | -0.203 | -0.111 | -0.018 | 0.109 | -0.136 | 0.084 |
| 24 | 18.59 | 22 | 0.6703 | -0.049 | -0.190 | 0.056 | 0.026 | -0.018 | -0.038 |

拟合模型后残差滞后各阶的白噪声检验结果均与 0 的差异无统计学意义,可以认为残差序列为白噪声,即所建立的模型已经充分概括了原始时间序列中蕴含的信息。

Autoregressive Factors

Factor 1: $1 + 0.99449 B + 0.43437 B^2$

综合前述考虑,应该拟合的模型为 ARIMA(2,1,0),由上结果可以得出 $\hat{\varphi}_1 = -0.99449$, $\hat{\varphi}_2 = -0.43437$,经假设检验(见 Conditional Least Squares Estimation 部分),两个估计值对应的 $P < 0.05$,均有统计学意义。

Forecasts for variable a1

| Obs | Forecast | Std Error | 95% Confidence Limits | |
|-----|-----------|-----------|-----------------------|-----------|
| 37 | 3706406.6 | 817.13 | 3704805.1 | 3708008.2 |
| 38 | 3737669.4 | 817.14 | 3736067.8 | 3739271.0 |
| 39 | 3768911.5 | 936.60 | 3767075.8 | 3770747.2 |
| 40 | 3800536.4 | 1003.40 | 3798569.8 | 3802503.0 |
| 41 | 3831789.6 | 1036.61 | 3829757.9 | 3833821.3 |
| 42 | 3863246.2 | 1111.79 | 3861067.1 | 3865425.2 |

这是对此医院 2006—2011 年间各年收入序列的预测结果。

27.4 谱 分 析

之前的各节一直是从时间域的角度进行分析的,探究时间数据之间的内在关联性特点。本节将从另一个全新的角度——频率域,来研究时间序列数据的内部规律性。时域(时间域)——自变量是时间,时域分析是用来描述响应变量随着时间变化的规律和特点的。频域(频率域)——自变量是频率,即时间的倒数,相应变量是通常说的频谱图。功率谱的意义是把单个时间函数的总能量分配到各个频率的振动上。频域分析的主要任务是揭示时间序列数据的周期性特征。

27.4.1 谱分析简介

应用时间序列分析的目的不外乎进行预测和控制。时域分析是通过建立时间序列模型对

时间数据样本进行预测和估计，展现数据内在的特性。频域数据则是从频率角度展现时间序列数据的特点和规律，其中最主要的任务是通过谱分析来获得时间数据的周期性特点，这一特点对于了解数据变化的规律来说是一个关键点。

时间序列研究对数据的要求是比较高的。最关键的是，数据必须是平稳序列的。首先要进行时间序列分析，对序列的长度是有要求的，长度不能太短，应该大样本；但是也不能太长，至少是周期的 2 倍以上。当然，时间序列的频域研究同样也要求测量的时间间隔应是等间距。另外，欲表达时间序列中周期值为 T 的信息成分，则采样间隔不能大于 $T/2$ ，该采样定理就是 Nyquist 采样定理。在进行谱分析之前要先进行数据的去趋势化，可以通过回归过程求剩余残差实现，也可以通过选项 ADJMEAN 来实现，本章会具体讲解该选项的具体内容。SAS 软件的 SPECTRA 过程中是不允许缺失数据出现的，缺失数据将无法参与分析，被自动排除于分析之外。在分析变量中出现缺失值时，程序自作主张将变量中没有缺失值的最长的连续数据部分作为分析变量。所以在分析时要注意检查缺失数据，也可以通过补充缺失数据来实现。

27.4.2 应用实例

【例 27-3】 太阳黑子运动的年度数据，该数据收集了 1749—1924 年的太阳黑子数数据（数据见表 27-3）。

表 27-3 Wolfer’s 太阳黑子运动数据

| | | | | | | | | | | | | | | | |
|-----|-----|-----|-----|------|------|-----|-----|------|-----|-----|-----|-----|------|------|-----|
| 809 | 834 | 477 | 478 | 307 | 122 | 96 | 102 | 324 | 476 | 540 | 629 | 859 | 612 | 451 | 364 |
| 209 | 114 | 378 | 698 | 1061 | 1008 | 816 | 665 | 348 | 306 | 70 | 198 | 925 | 1544 | 1259 | 848 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 439 | 186 | 57 | 36 | 14 | 96 | 474 | 571 | 1039 | 806 | 636 | 376 | 261 | 142 | 58 | 167 |

27.4.3 SAS 程序

SAS 程序名为 sastjfx27_3.sas。

```
data sunpot;
input wolfer@@ ;
year=intnx('year','1jan1749'd,
           _n_-1);
format year year4.;
cards;
809 834 477478 307 122 96
102 324 476540 629 859 612
.....
1039 806 636376 261 14258
167
;
run;
ods html;
symbol i=spline v=plus;
proc gplot data=sunpot;

plot wolfer* year;
run;
proc spectra data=sunpot out=spec p
ph s coef adjmean whitetest;
var wolfer;
weights1 2 3 2 1;
run;
proc gplot data=spec;
plot p_01* period;
plot s_01* period;
run;
ods html close;
```

【程序说明】 gplot 过程步通过绘制散点图先从整体上了解数据是否存在一定的周期变化规律，初步了解数据变化特性。spectra 过程步用来进行序列的谱分析，通过该过程将原始的

时域数据转换为频域数据,并将获得的频域数据储存在新的数据集 spec 中,该步中可以同时进行白噪声检验。检验该序列是否确实存在一定的趋势或者规律而不是一个随机波的序列。proc spectra 语句中使用 p、ph、s、coef 选项,将周期图变量、交叉谱的相位变量、谱密度估计和周期图、傅里叶系数等写入数据集 spec 中,adjmean 选项使频率为 0 的值为 0,从而避免了将那点排除在散点图外,whitetest 选项表示输出对序列是白噪声假设的检验。

27.4.4 主要分析结果及解释

图 27-2 给出两个重要提示:(1)太阳黑子运动序列随着时间的变化均数基本处于一个水平的位置,没有太大的变化,方差似乎也没有随着时间变化,基本保持一个水平,即该数据满足平稳序列的要求;(2)太阳黑子运动每隔 10 年左右就会出现一个峰值,这是原始数据给出的提示,但是 10 年是不是太阳黑子运动的一个周期,还需要做进一步的分析来证实。本例中数据满足了平稳序列的要求,但是多数数据是不满足该条件的,可以先通过差分等方法实现数据的平稳化,再做进一步的分析。

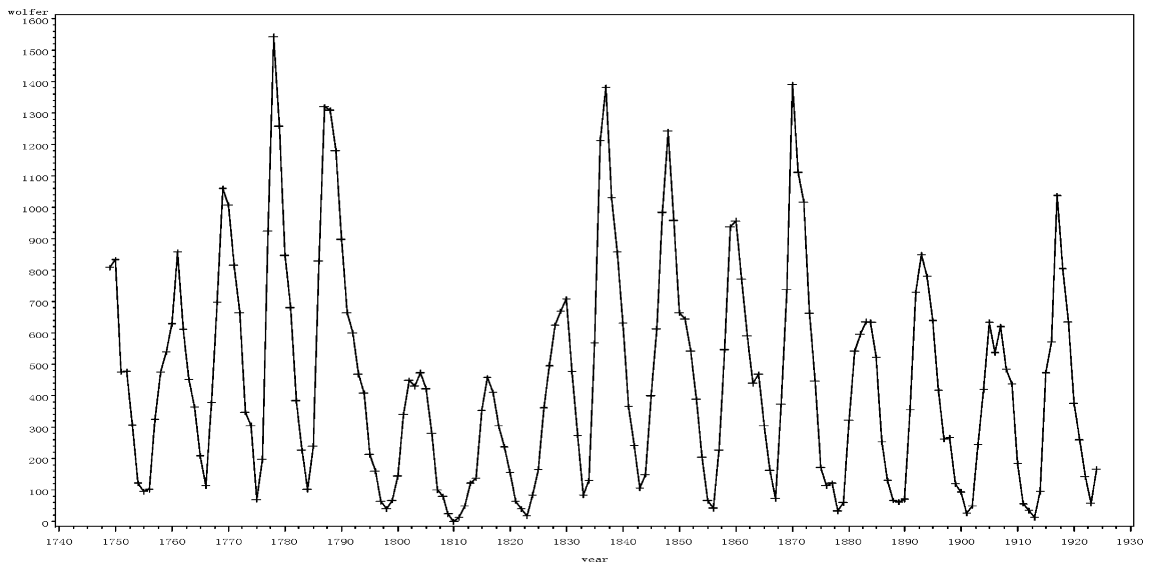


图 27-2 太阳黑子运动序列图

```

The SPECTRA Procedure
Test for White Noise for
Variable wolfer

M - 1                87
Max( P( * ) )       4062267
Sum( P( * ) )       21156512

Fisher's Kappa: (M - 1) * Max( P( * ) ) / Sum( P( * ) )
Kappa                16.70489

Bartlett's Kolmogorov-Smirnov Statistic;
Maximum absolute difference of the standardized
partial sums of the periodogram and the CDF of a
uniform(0,1) random variable.

```

| | |
|---------------------|----------|
| Test Statistic | 0.650055 |
| Approximate P-Value | < .0001 |

这是白噪声的检验结果。

图 27-3 是周期与周期图的散点图。可以这样理解周期图：各个周期出现的强度，在这个强度远远超过其他的周期的强度时，就可以认为该值是该序列的一个周期值。由图 27-3 发现图中有多个峰，哪个才是该序列的周期值呢？很难界定，但是由于周期图不是功率谱的一致估计，具体表现在图上会有很多峰，并呈现剧烈的抖动，这就需要通过谱窗来进行平滑（平滑的原理是赋予各个时间点不同的权重，距离越远，影响越小）。平滑后的数据是 S（谱密度估计值）。通过谱密度与周期做散点图（见图 27-4），可以发现之前的抖动现象已经得到改善。呈现一枝独秀的景象，那么随之可以确定周期值，谱密度最高值作为周期值，即图中的峰值对应的周期值。由此可以得出结论：太阳黑子运动的周期是 11 年，印证了第一步中的想法。

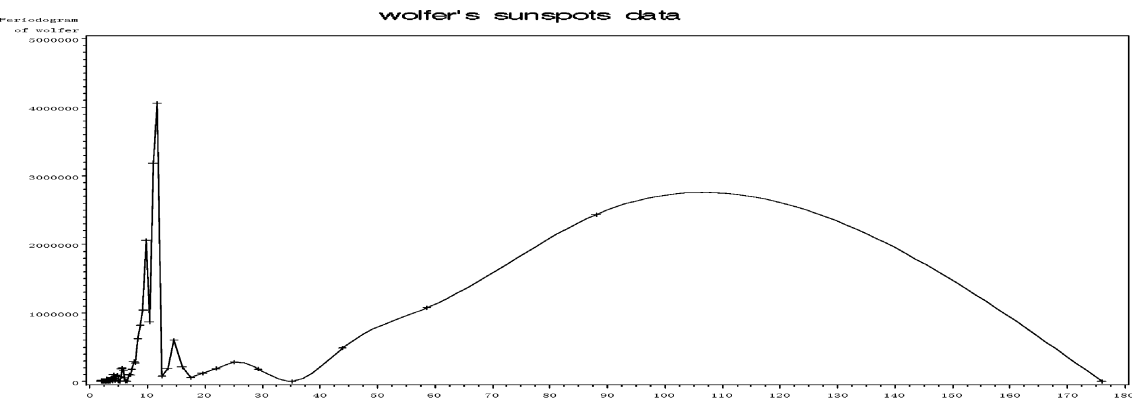


图 27-3 太阳黑子运动数周期-周期图散点图

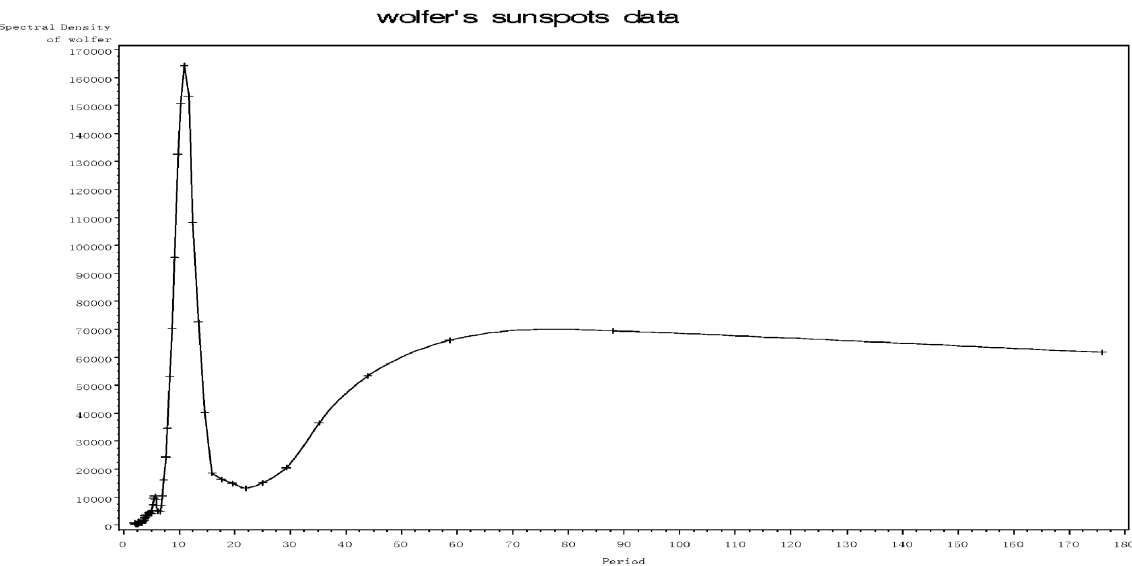


图 27-4 太阳黑子运动数周期 - 谱密度估计值散点图

27.5 X12 方 法

27.5.1 X12 过程简介

X12 过程是根据美国人口普查局 X-12-ARIMA 季节调整程序改编的,用于调整月度或季度时间序列数据。该过程包含了 X-11 过程(Shiskin, Young, and Musgrave 1967)、X-11-ARIMA/88 模型(Dagum 1988)以及一些新的特征(Findley 1998)。X12 过程较 X11 的一个主要提高是应用 regARIMA 模型——带有 ARIMA(Autoregressive Integrated Moving Average)误差的回归模型,利用该模型进行移动假日、月份长度、交易日效应等固定效应的调整。X-12-ARIMA 模型包含了美国人口统计局和加拿大统计局开发的季节调整模型的主要特征。

对序列进行季节调整是基于这样的假定:季节性波动可以由原始序列($Q_t, t=1, \dots, n$)中测得,并能与趋势起伏、交易日及不规则波动分离开;这一时间序列的季节成分(S_t)定义为年内的变动,从一年到一年之间恒定地取值或缓慢地变化;趋势起伏项(C_t)包含了由长期趋势,经济起伏及其他长期起伏因素引起的变化;交易日成分(D_t)是由日历中交易日位置变化引起的;不规则成分(I_t)是残余的变化量。

27.5.2 应用实例

【例 27-4】 已知某零售企业 1994 年 1 月—2005 年 12 月的零售额数据,考虑到季节因素或交易日因素对序列观测值的影响,需要对其进行季节调整(数据见表 27-4)。

表 27-4 某零售企业 1994 年 1 月—2005 年 12 月的零售额数据

| | | | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 115 | 126 | 141 | 135 | 125 | 149 | 170 | 170 | 158 | 133 | 114 | 140 |
| 145 | 150 | 178 | 163 | 172 | 178 | 199 | 199 | 184 | 162 | 146 | 166 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 417 | 391 | 419 | 461 | 472 | 535 | 622 | 606 | 508 | 461 | 390 | 432 |

27.5.3 SAS 程序

不考虑交易日因素,进行单纯的季节调整时,所用 SAS 程序名为 sastjfx27_4A.sas。

```

data sales;
  set sashelp.air;
  sales = air;
  date = intnx('month', '01jan94'd,
    _n_ - 1);
  format date monyy.;
run;
proc x12 data = sales date = date;
  var sales;
  transform function = auto;
run;
ods html;
ods graphics on;
proc x12 data = sales date = date;

ods graphics off;
ods html close;
proc x12 data = sales date = date;
  var sales;
  transform power = 0;
  arima model = ((0,1,1) (0,1,1));
  estimate;
run;
proc x12 data = sales date = date;
  var sales;
  transform power = 0;
  arima model = ((0,1,1) (0,1,1));
  estimate;
  x11;

```

```
var sales;
transform power = 0;
identify diff = (0,1)
sdiff = (0,1);
automdl;
run;
```

【程序说明】 数据步调用 SASHELP 逻辑库中的 air 数据集，并重新命名为 sales，同时指定 sales 的时间范围。第一次调用 X12 过程是利用 TRANSFORM 自动识别序列是否需要转换，transform 语句是对输入的数据集进行适当转换，便于估计合适的 regARIMA 模型，“function = auto”表示基于 AIC 准则决定是否进行 LOG 转换，“date = date”指定存有每个观测的日期的变量。第二次调用 X12 过程中，“transform power = 0;”表示对输入的数据集进行对数变换，采用 IDENTIFY 语句通过 ACF 与 PACF 以及 AUTODML 的自动过程进行模型识别，综合判断选择合适的模型。根据前两个过程得出的结果，确定了最终选择的 ARIMA 模型是(0 1 1)(0 1 1)₁₂，接着用 ARIMA 和 ESTIMATE 语句替换 IDENTIFY 语句。结果输出了参数估计和估计的结论统计量。最后一次调用 X12 过程，是在模型是合适的情况下，通过 X11 语句进行季节调整。

考虑交易日因素，进行季节调整时，所用 SAS 程序名为 sastjfx27_4B.sas。

```
data sales;
set sashelp.air;
sales = air;
date = intnx('month', '01jan94'd,
            _n_ - 1);
format date monyy.;
run;
ods html;
proc x12 data = sales date = date;
var sales;

transform function = log;
regression predefined = td;
automdl print = unitroottestmdl
best5model;

estimate;
x11;
output out = out a1 a2 a6 b1 c17
c20 d1 d7 d8 d9 d10 d11 d12
d13 d16 d18;
run;
ods html close;
```

【程序说明】 “regression predefined = td;”语句指定 regARIMA 模型或需要在 identify 语句中处理的特定回归变量为交易日效应。“print = unitroottestmdl best5model”打印使用 TRAMO 方法进行 ARIMA 模型识别过程中的单位根识别结果和按照 TRAMO 自动模型识别方法自动识别后排在 前 5 位的 ARIMA 模型。“estimate”用于估计 REGRESSION 与 ARIMA 语句指定的 regARIMA 模型。“output out = out a1 a2 a6 b1 c17 c20 d1 d7 d8 d9 d10 d11 d12 d13 d16 d18;”语句是将一些统计量输出到数据集 out 中，这些统计量分别是原始序列、预调整因子、regARIMA 交易日成分、预调整或原始序列、不规则成分的最终权重、最终极端值调整因子、修饰的原始序列(D 迭代)、初步趋势循环(D 迭代)、最终的未调整的 S-I 比率、极端 S-I 比率的最终替换值、最终季节因子、最终季节调整序列、最终趋势循环、最终不规则序列、联合调整因子、联合日历调整因子。

27.5.4 摘录主要分析结果

以下是未考虑交易日因素、进行单纯的季节调整时的结果。

Results of Automatic Transformation Selection
For variable sales
AICC(with aicdiff = -2.00) prefers Log(sales)
Adjustment will be Multiplicative

这是自动数据转换形式的选择结果, 可知根据 AICC 的值选择合适的数据转换形式为 LOG 转换。

图 27-5 ~ 图 27-8 是模型识别过程的自相关和偏自相关函数图。由此可见, $\text{diff} = 1$, $\text{sdiff} = 1$ 时 ACF、PACF 图符合平稳性条件, 且 PACF 缓慢衰减并趋于 0, 而 ACF 显示一阶后截尾, 及滞后 12 阶时与零有显著性差异的现象, 故初步断定合适的模型为 $(0, 1, 1)(0, 1, 1)_{12}$ 。

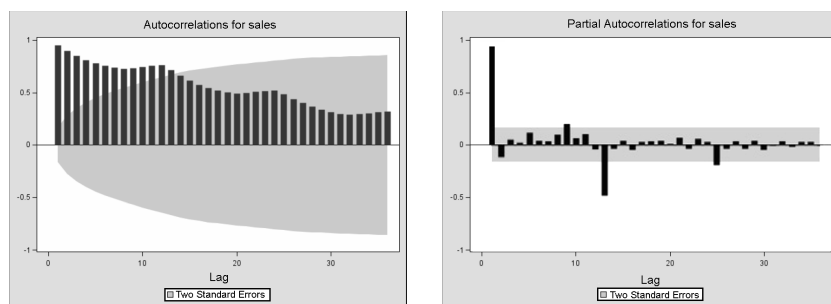


图 27-5 $\text{diff} = 0$ $\text{sdiff} = 0$ 时的 ACF、PACF

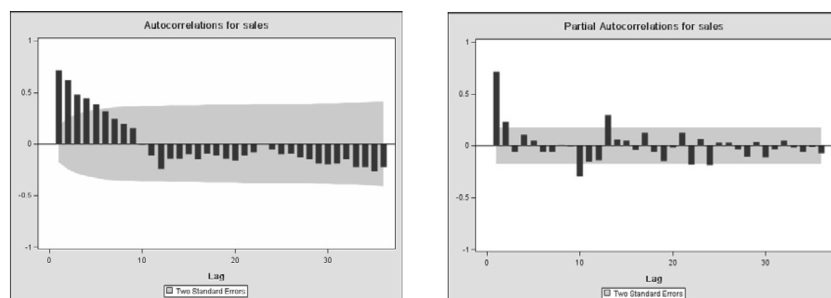


图 27-6 $\text{diff} = 0$ $\text{sdiff} = 1$ 时的 ACF、PACF

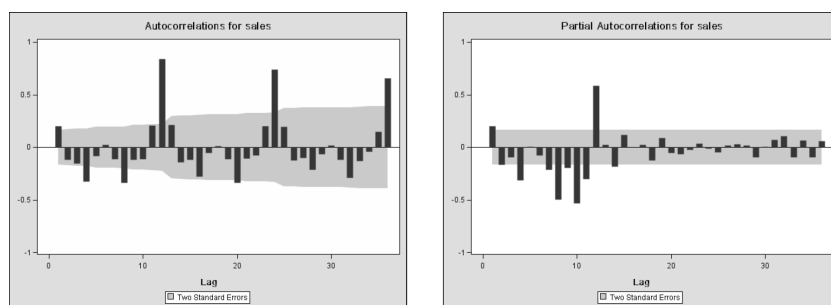


图 27-7 $\text{diff} = 1$ $\text{sdiff} = 0$ 时的 ACF、PACF

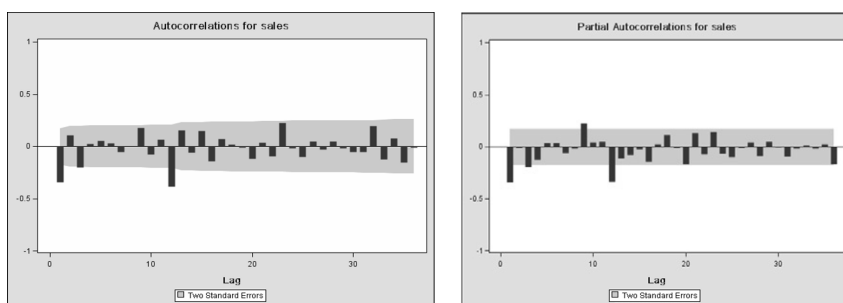


图 27-8 $\text{diff} = 1$ $\text{sdiff} = 1$ 时的 ACF、PACF

| Final Automatic Model Selection | |
|---|-----------------|
| For variable sales | |
| Source of Model | Estimated Model |
| Automatic Model Choice(0, 1, 1) (0, 1, 1) | |

由上结果可知，自动模型识别的最终模型为(0, 1, 1)(0, 1, 1)，与 ACF、PACF 图识别结果一致。

| Exact ARMA Maximum Likelihood Estimation | | | | | |
|--|-----|----------|----------------|---------|---------|
| For variable sales | | | | | |
| Parameter | Lag | Estimate | Standard Error | t Value | Pr > t |
| Nonseasonal MA | 1 | 0.40181 | 0.07887 | 5.09 | <.0001 |
| Seasonal MA | 12 | 0.55695 | 0.07626 | 7.30 | <.0001 |

非季节和季节 MA 检验结果均为拒绝无效假设，可认为所选模型是合适的。

| Table D 8. A: F-tests for Seasonality For variable sales | | | | | |
|--|----------------|-----|-------------|----------|----|
| Test for the presence of seasonality assuming stability. | | | | | |
| | Sum of Squares | DF | Mean Square | F-Value | |
| Between Months | 23571.17 | 11 | 2142.833 | 190.7143 | ** |
| Residual | 1483.13 | 132 | 11.23583 | | |
| Total | 25054.3 | 143 | | | |
| ** Seasonality present at the 0.1 per cent level. | | | | | |

| Nonparametric Test for the Presence of Seasonality Assuming Stability | | |
|---|----|-------------------|
| Kruskal-Wallis Statistic | DF | Probability Level |
| 131.856 | 11 | .00% |
| Seasonality present at the one percent level. | | |

| Moving Seasonality Test | | | | |
|--|----------------|-----|-------------|----------|
| | Sum of Squares | DF | Mean Square | F-Value |
| Between Years | 251.4713 | 11 | 22.86103 | 2.873598 |
| Error | 962.6205 | 121 | 7.955541 | * |
| * Moving seasonality present at the one percent level. | | | | |

| COMBINED TEST FOR THE PRESENCE OF IDENTIFIABLE SEASONALITY | |
|--|---------|
| IDENTIFIABLE SEASONALITY | PRESENT |

以上 4 部分结果都属于 Table D8. A 季节性 F-检验的一部分，通过以上结果可以看出，该序列中存在有统计学意义的季节因素，所以需要调整。

以下是考虑交易日因素、进行季节调整时的结果。

| ARIMA Estimates for Unit Root Identification | | | | | |
|--|-------------------|-----------------|-----------|-----------|----------|
| For variable sales | | | | | |
| Model Number | Estimation Method | Estimated Model | | ARMA | |
| | | | | Parameter | Estimate |
| 1 | H-R | (2, 0, 0) | (1, 0, 0) | NS_AR_1 | 0.67979 |

| | | | | | |
|---|-----|-----------|-----------|---------|----------|
| 2 | H-R | (2, 0, 0) | (1, 0, 0) | NS_AR_2 | 0.27814 |
| | H-R | (2, 0, 0) | (1, 0, 0) | S_AR_12 | 0.91022 |
| | H-R | (1, 1, 1) | (1, 0, 1) | NS_AR_1 | -0.29340 |
| | H-R | (1, 1, 1) | (1, 0, 1) | S_AR_12 | 0.99422 |
| 3 | H-R | (1, 1, 1) | (1, 0, 1) | NS_MA_1 | -0.06303 |
| | H-R | (1, 1, 1) | (1, 0, 1) | S_MA_12 | 0.57326 |
| | H-R | (1, 1, 1) | (1, 1, 1) | NS_AR_1 | -0.21870 |
| | H-R | (1, 1, 1) | (1, 1, 1) | S_AR_12 | -0.25561 |
| | H-R | (1, 1, 1) | (1, 1, 1) | NS_MA_1 | 0.02777 |
| | H-R | (1, 1, 1) | (1, 1, 1) | S_MA_12 | 0.44680 |

这部分结果列出了使用 TRAMO 方法进行 ARIMA 模型识别过程中的单位根识别结果。

Best Five ARIMA Models Chosen by Automatic

Modeling

For variable sales

| Rank | Estimated Model | | BIC2 |
|------|-----------------|-----------|----------|
| 1 | (0, 1, 1) | (1, 1, 1) | -3.74241 |
| 2 | (0, 1, 1) | (0, 1, 1) | -3.73673 |
| 3 | (1, 1, 0) | (1, 1, 1) | -3.73360 |
| 4 | (1, 1, 1) | (1, 1, 1) | -3.72111 |
| 5 | (0, 1, 2) | (1, 1, 1) | -3.71131 |

Final Automatic Model Selection

For variable sales

| Source of Model | Estimated Model | |
|------------------------|-----------------|-----------|
| Automatic Model Choice | (0, 1, 1) | (1, 1, 1) |

这部分结果列出了使用 TRAMO 方法进行自动模型识别 regARIMA 模型的过程中列为前五位的模型以及最终识别的 regARIMA 模型形式。

Regression Model Parameter Estimates

For variable sales

| Type | Parameter | Estimate | Standard Error | t Value | Pr > t |
|-------------|----------------|----------|----------------|---------|---------|
| Trading Day | MON | 0.00677 | 0.00477 | 1.42 | 0.1580 |
| | TUE | -0.01526 | 0.00484 | -3.16 | 0.0020 |
| | WED | 0.00701 | 0.00493 | 1.42 | 0.1577 |
| | THU | -0.01085 | 0.00491 | -2.21 | 0.0290 |
| | FRI | 0.00644 | 0.00503 | 1.28 | 0.2031 |
| | SAT | 0.00464 | 0.00503 | 0.92 | 0.3582 |
| | SUN(derived) * | 0.00125 | 0.00474 | 0.26 | 0.7919 |

Chi-squared Tests for Groups of Regressors

For variable sales

| Regression Effect | DF | Chi-Square | Pr > ChiSq |
|-------------------|----|------------|------------|
| Trading Day | 6 | 35.1299 | <.0001 |

这里首先列出了各交易日因子的检验结果，其中周二、周四有统计学意义。然后又给出总的结论，即交易日效应是存在的。

27.6 本章小结

本章通过多个例子介绍了时间序列数据的处理方法，特别是结合 SAS 软件中的有关过程，重点介绍了以下主要内容：指数平滑法、ARIMA 过程(求和自回归滑动平均过程)、谱分析和 X12 季节调整过程及其具体使用方法。

(张晋昕 杨业春 薛允莲 张熙 王霄 刘明华 李长平 高辉)

第 6 篇 对定性结果进行预测性分析

第 28 章 非配对设计定性资料多重 logistic 回归分析

生物医学研究中最常见的问题之一是探索各种影响因素(自变量 X)与疾病或健康(响应变量 Y)之间的关系。许多情况下,疾病和健康状况属于分类变量,包括二值变量、多值有序变量和多值名义变量。当响应变量为分类变量时,就不适合使用线性回归进行分析,这时可以考虑采用多重 logistic 回归。本章将结合实例介绍如何用 SAS 实现非配对设计的多重 logistic 回归分析。

28.1 问题、数据及统计分析方法的选择

28.1.1 问题与数据

【例 28-1】 为探讨儿童伯基特淋巴瘤与疟疾感染和 EB 病毒感染之间的关系,2005—2006 年,在非洲马拉维的儿童肿瘤中心选取了 126 例伯基特淋巴瘤患者(≤ 15 岁)作为病例组,在同一医院选择了 89 例患有恶性非血液系统肿瘤或其他良性疾病的患者(≤ 15 岁)作为对照组。检测患者的疟疾抗体和 EB 病毒抗体,疟疾抗体检测的 OD 值小于 0.2 判定为阴性或低浓度抗体,大于等于 0.2 判定为中高浓度抗体;EB 病毒抗体滴度小于等于 1:640 判定为低浓度抗体,大于等于 1:1280 判定为中高浓度抗体。有关研究资料的总结见表 28-1。

表 28-1 肿瘤中心患者 EBV 和疟疾感染情况

| 疟疾抗体高低 (X_1) | EB 病毒抗体滴度 (X_2) | 例数(人) | | |
|---------------------|------------------------|---------------|----|-----|
| | | 是否患伯基特淋巴瘤(Y): | 患病 | 未患病 |
| 阴性/低 | 低 | | 5 | 22 |
| | 中/高 | | 32 | 25 |
| 中/高 | 低 | | 7 | 16 |
| | 中/高 | | 82 | 26 |

【例 28-2】 为比较两种药物(A 药和 B 药)治疗关节扭伤的疗效,进行多中心(4 家医院)、随机、对照、双盲、双模拟临床试验,将受试者随机分入不同的处理组,观察不同处理方法的疗效。有关资料总结见表 28-2。

表 28-2 两种药物治疗关节扭伤的疗效

| 药 | 物 | 医 | 院 | 例数(人) | | | | |
|-----|---|---|---|-------|----|----|----|----|
| | | | | 疗效: | 痊愈 | 显效 | 好转 | 无效 |
| A 药 | | 1 | | 6 | 16 | 5 | 3 | |
| | | 2 | | 8 | 12 | 6 | 4 | |
| | | 3 | | 8 | 15 | 5 | 2 | |
| | | 4 | | 5 | 8 | 12 | 3 | |
| B 药 | | 1 | | 5 | 6 | 12 | 6 | |
| | | 2 | | 4 | 9 | 8 | 7 | |
| | | 3 | | 6 | 10 | 9 | 5 | |
| | | 4 | | 4 | 11 | 8 | 5 | |

【例 28-3】 在某社区随机抽取了 100 名育龄妇女进行问卷调查,了解她们的年龄(A)、生育史(B)、收入(C)及采取的避孕方式(Y),评价哪些因素会影响育龄妇女对避孕方式的选择。将调查对象按年龄分为小于 30 岁组(A=1),30~35 岁组(A=2),35 岁以上组(A=3);按生育史分为未生育组(B=0)和已生育组(B=1);按收入水平分为低收入组(C=1)和高收入组(C=2);选择的避孕方式有 IUD(Y=1)、口服避孕药(Y=2)和避孕套(Y=3)3 种。有关资料见表 28-3。

表 28-3 育龄女性选择避孕方式的影响因素

| 编号 | 年龄(A) | 生育史(B) | 收入(C) | 避孕方法(Y) | 编号 | 年龄(A) | 生育史(B) | 收入(C) | 避孕方法(Y) |
|-----|-------|--------|-------|---------|-----|-------|--------|-------|---------|
| 1 | 2 | 1 | 2 | 2 | 51 | 2 | 1 | 1 | 1 |
| 2 | 2 | 1 | 1 | 1 | 52 | 3 | 0 | 2 | 3 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 50 | 2 | 1 | 2 | 2 | 100 | 1 | 0 | 2 | 3 |

28.1.2 对数据结构的分析

对例 28-1 而言,该研究设计类型属于非配对病例对照调查设计,属于成组设计,研究目的是分析感染疟疾(自变量 X_1)、感染 EB 病毒(自变量 X_2)是否是罹患伯基特淋巴瘤(响应变量 Y)的危险因素。研究对象为儿童肿瘤中心的患者,按照是否罹患伯基特淋巴瘤分为病例组和对照组;研究因素为疟疾、EB 病毒感染状况;研究效应为是否罹患伯基特淋巴瘤。

这是响应变量为二值变量的三维列联表资料,有 1 个二值响应变量(是否患伯基特淋巴瘤: $Y=1$, 患此病; $Y=0$, 未患此病); 2 个二值自变量(疟疾抗体滴度高低: $X_1=1$ (高)、 0 (低); EB 病毒抗体滴度高低: $X_2=1$ (高)、 0 (低))。

对例 28-2 而言,此研究的目的是比较两种药物治疗同一种疾病的疗效,属于临床试验研究。试验对象是患有关节扭伤的病人,试验因素是药物(A 药和 B 药)和医院(4 家医院),试验效应是治疗效果。

表 28-2 属于结果变量为多值有序变量的三维列联表资料,响应变量为疗效,是多值有序变量($Y=1\sim4$),分别对应痊愈、显效、好转、无效 4 个水平。自变量有 2 个: 1 个为二值变量药物种类,分别是 A 药($X_1=1$)和 B 药($X_1=0$);另一自变量为多值名义变量医院($X_2=1\sim4$),分别对应 4 家不同医院。

对例 28-3 而言,此研究的目的是探索影响育龄妇女选择避孕方式的因素,属于横断面调查研究。研究对象是某社区育龄妇女,研究因素是年龄、生育史和收入,研究效应是对避孕方法的选择。

研究中有 1 个响应变量(避孕方式),它有 3 个取值,分别为 IUD($Y=1$)、口服避孕药($Y=2$)和避孕套($Y=3$),这 3 个取值之间无内在的有序关联,即响应变量为多值名义变量。研究中还有 3 个自变量:2 个是二值变量(生育史和收入),1 个是多值有序变量(年龄)。本例中是以数据库形式呈现资料,也可以将其整理成高维列联表形式。

28.1.3 分析目的与统计分析方法的选择

就例 28-1 而言,对该资料进行分析可以采用 CMH χ^2 检验、对数线性模型或 logistic 回归分析。若是进行差异性分析,可以选择 CMH χ^2 检验;若是进行探索性研究,除分析各自变量对因变量的影响外,还要考察各个自变量间的关系以及各自变量是否对因变量有联合作用,可选择对数线性模型分析;若是定量地分析自变量对结果变量发生概率的影响,特别是在对发病的优势比或相对危险度进行评价时,可以选择 logistic 回归分析。本章采用 logistic 回归对该资料进行分析。

就例 28-2 而言,对该资料进行分析可以采用 CMH 校正的秩和检验或有序变量的多重 logistic 回归分析。如果要在控制其他影响因素的基础上,比较某一个影响因素不同水平间的疗效是否存在差异,可以采用 CMH 校正的秩和检验;如果想定量地分析影响因素对结果变量发生概率的影响,可以选择进行 logistic 回归。本章采用累积 logistic 回归模型对该资料进行分析。

就例 28-3 而言,分析例 28-3 中的资料,也就是结果变量为多值名义变量的高维列联表资料时,可以选用的统计分析方法有 CMH χ^2 检验、扩展的多重 logistic 回归分析和对数线性模型。这里使用扩展的多重 logistic 回归分析该资料,此模型也叫做多项 logit 模型。

28.1.4 logistic 回归分析简介

logistic 回归属于概率型回归,其应用范围很广,不仅适用于流行病学上病因学的分析,也可用于临床疗效评价、卫生服务研究等。它适用于因变量是定性变量的情形,对自变量的数目、性质没有特殊要求。

按照因变量的类型可以将 logistic 回归分为三类:因变量为二值变量的 logistic 回归;因变量为多值有序变量的 logistic 回归,称为累积 logistic 回归模型或序次 logistic 回归模型;因变量为多值名义变量的 logistic 回归,称为多项 logit 模型。按照设计类型可以将 logistic 回归模型分为非条件 logistic 回归和条件 logistic 回归,其中非条件 logistic 回归是指一般的 logistic 模型,适用于成组设计资料;条件 logistic 回归则是针对配对设计资料。

28.2 二值变量的多重 logistic 回归分析

28.2.1 SAS 程序及说明

设对例 28-1 中资料进行 logistic 回归分析的程序名为 SASTJFX28_1.SAS,程序如下:

```
data SASTJFX28_1;
  input Y X1 X2 count @@ ;
  cards;
1 0 0 5 1 0 1 32
1 1 0 7 1 1 1 82
0 0 0 22 0 0 1 25
0 1 0 16 0 1 1 26
;
run;

ODS HTML STYLE = MINIMAL;
proc logistic data = SASTJFX28_1
  descending;
  model Y = X1 X2 / rsq cl stb
  scale = none aggregate;
  freq count;
run;
ODS HTML CLOSE;
```

【程序说明】 首先建立数据集 prg28_1， X_1 、 X_2 、 Y 和 count 分别表示疟疾抗体浓度高低、EB 病毒抗体滴度、是否患伯基特淋巴瘤和频数。语句“proc logistic”表示调用 LOGISTIC 过程，该语句中的“descending”选项要求按照降序输出结果，即输出 Y 取值较大时 ($Y = 1$) 的参数估计值；MODEL 语句中的选项“rsq”和“cl”分别要求输出广义决定系数和回归系数的 95% 置信区间；“stb”选项要求输出标准化回归系数；“scale = none aggregate”选项要求对模型进行拟合优度检验。由于资料以列联表的形式给出，所以还使用语句“freq count;”来指定频数变量为 count。

28.2.2 主要分析结果及结论

例 28-1 资料的主要输出结果如下：

| The LOGISTIC Procedure | |
|---------------------------|------------------|
| Model Information | |
| Data Set | WORK. PRG28_1 |
| Response Variable | Y |
| Number of Response Levels | 2 |
| Frequency Variable | count |
| Model | binary logit |
| Optimization Technique | Fisher's scoring |

| Response Profile | | |
|------------------|---|-----------------|
| Ordered Value | Y | Total Frequency |
| 1 | 1 | 126 |
| 2 | 0 | 89 |

Probability modeled is $Y = 1$.

结果第一部分输出模型基本信息，其中数据集名称为 prg28_1，响应变量为 Y ，它有两个水平，频数变量为 count，使用的模型是二值 logit 模型，参数估计时的优化方法是 Fisher's scoring 法。响应变量的取值顺序为 1 和 0，各自分别有 126 例和 89 例。最后一行文字说明该模型以 1 为基础，也就是以患病的概率为基础建模。这里需要注意的是，在 LOGISTIC 过程中，默认状态下是以响应变量取值较小的那个水平的发生概率为基础建模。如果原始数据集中是以 1 来表示阳性结果，以 0 表示阴性结果的话，最好在 proc logistic 语句之后使用 descending 选项来改变响应变量的取值顺序，这样得到的方程可以直接计算阳性事件发生的概率。

Deviance and Pearson Goodness-of-Fit Statistics

| Criterion | Value | DF | Value/DF | Pr > ChiSq |
|-----------|--------|----|----------|------------|
| Deviance | 0.1058 | 1 | 0.1058 | 0.7450 |
| Pearson | 0.1065 | 1 | 0.1065 | 0.7442 |

| Model Fit Statistics | | |
|----------------------|----------------|--------------------------|
| Criterion | Intercept Only | Intercept and Covariates |
| AIC | 293.654 | 257.622 |
| SC | 297.025 | 267.734 |
| -2 Log L | 291.654 | 251.622 |

| | | | |
|----------|--------|-------------------------|--------|
| R-Square | 0.1699 | Max - rescaled R-Square | 0.2288 |
|----------|--------|-------------------------|--------|

Testing Global Null Hypothesis: BETA = 0

| Test | Chi-Square | DF | Pr > ChiSq |
|------------------|------------|----|------------|
| Likelihood Ratio | 40.0318 | 2 | < .0001 |
| Score | 38.6691 | 2 | < .0001 |
| Wald | 32.3410 | 2 | < .0001 |

第二部分输出拟合优度检验以及对整个模型进行假设检验的结果。SAS 提供了两种方法的拟合优度检验，P 值分别为 0.7450 和 0.7442，均提示模型对资料总体上拟合效果较好。在“Model Fit Statistics”部分输出了三种不同的模型拟合统计量，分别是 AIC、SC 和 -2LogL ，这三个统计量取值越小，说明模型拟合越好。结果表明在包含自变量的模型中，这三个统计量的取值都小于不包含自变量的模型。另外，还输出了广义决定系数，广义决定系数与线性回归中的决定系数类似，其取值为 0.1699，提示总变异中有 16.99% 可由自变量的变异来解释，自变量能够很好地解释响应变量的变化。在“Testing Global Null Hypothesis: BETA = 0”部分列出了对整个模型进行假设检验的结果，它的原假设是所有的回归系数都为 0，分别使用似然比检验、计分检验和 Wald 检验三种方法。检验结果中依次给出了 χ^2 值、自由度和 P 值，可以看出，三种方法的 P 值都小于 0.0001，可以认为该模型是成立的。

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq | Standardized Estimate |
|-----------|----|----------|----------------|-----------------|------------|-----------------------|
| Intercept | 1 | -1.5901 | 0.3802 | 17.4938 | < .0001 | |
| X1 | 1 | 0.8497 | 0.3103 | 7.4990 | 0.0062 | 0.2291 |
| X2 | 1 | 1.8675 | 0.3779 | 24.4251 | < .0001 | 0.4360 |

第三部分是参数估计的结果，由左至右的各列依次为变量名、自由度、参数估计值、标准误、对各参数进行检验的 Wald χ^2 值和 P 值以及标准化回归系数。本资料中自变量 X_1 和 X_2 对应的 P 值分别为 0.0062 和小于 0.0001，提示两个自变量的回归系数的估计值均有统计学意义。当 P 值大于等于 0.05 时，需剔除无统计学意义的自变量后再重新进行模型中参数估计值的计算。本例中可得到 logistic 回归模型

$$P(Y = 1) = \frac{e^{-1.5901 + 0.8497X_1 + 1.8675X_2}}{1 + e^{-1.5901 + 0.8497X_1 + 1.8675X_2}}$$

其中, P 代表罹患伯基特淋巴瘤的概率, $1 - P$ 则代表未患伯基特淋巴瘤的概率。使用 logistic 回归表达式可筛选伯基特淋巴瘤的危险因素, 由于 X_1 和 X_2 的回归系数均为正值, 因此感染疟疾或感染 EB 病毒都是罹患伯基特淋巴瘤的危险因素。注意, 若 SAS 程序语句中未使用 descending 选项, 则回归系数正值的含义相反。

这部分结果的最右侧一列输出了标准化回归系数。由于各个自变量的量纲不同, 因此回归系数不能直接进行比较, 而标准化回归系数则可以评价各自变量的作用大小, X_1 的标准化回归系数为 1.2668, X_2 的标准化回归系数为 2.4106, 提示感染疟疾对罹患伯基特淋巴瘤的作用不如感染 EB 病毒那样强。

| Odds Ratio Estimates | | | |
|----------------------|----------------|----------------------------|--------|
| Effect | Point Estimate | 95% Wald Confidence Limits | |
| X1 | 2.339 | 1.273 | 4.297 |
| X2 | 6.472 | 3.086 | 13.573 |

第四部分输出优势比, 包括其估计值和 95% 可信区间。 X_1 和 X_2 的 OR 值分别为 2.339 和 6.472, 95% 可信区间分别为 (1.273, 4.297) 和 (3.086, 13.573), 这两个可信区间都不包含 1, 说明 OR 值与 1 之间的差别有统计学意义。

| Association of Predicted Probabilities and Observed Responses | | | |
|---|-------|-----------|-------|
| Percent Concordant | 58.3 | Somers' D | 0.447 |
| Percent Discordant | 13.6 | Gamma | 0.622 |
| Percent Tied | 28.1 | Tau-a | 0.218 |
| Pairs | 11214 | c | 0.723 |

第五部分输出的是预测概率和观察响应之间的关联性, 需要注意的是统计量 c 的取值即为经常用到的 ROC 曲线的曲线下面积。

| Wald Confidence Interval for Parameters | | | |
|---|----------|-----------------------|---------|
| Parameter | Estimate | 95% Confidence Limits | |
| Intercept | -1.5901 | -2.3353 | -0.8450 |
| X1 | 0.8497 | 0.2416 | 1.4579 |
| X2 | 1.8675 | 1.1269 | 2.6081 |

第六部分输出了截距及回归系数的估计值及其 95% 可信区间, 截距和回归系数的 95% 可信区间均不包括 0, 也提示截距和回归系数不为 0。

专业结论: 是否罹患伯基特淋巴瘤与感染疟疾或感染 EB 病毒有关, 疟疾抗体浓度为中/高水平的儿童患伯基特淋巴瘤的风险是疟疾抗体为阴性/低浓度儿童的 2.339 倍; EB 病毒抗体滴度为中高浓度者患伯基特淋巴瘤的风险是低浓度者的 6.472 倍。

28.3 多值有序变量的多重 logistic 回归分析

28.3.1 SAS 程序及说明

设对例 28-2 中资料进行累积 logistic 回归分析的程序名为 SASTJFX28_2.SAS, 程序如下:

```

data SASTJFX28_2;
do X1 = to 2
do X2 = 1 to 4;
do Y = 1 to 4;
input count@@ ;
output;
end;
end;
end;
cards;
6 16 5 3 8 12 6 4 8 15 5 2
5 8 12 3 5 6 12 6 4 9 8 7
6 10 9 5 4 11 8 5
;
run;

ODS HTML STYLE = MINIMAL;
proc logistic data = SASTJFX28_2;
freq count;
class X2;
model Y = X1 X2;
run;
proc logistic data = SASTJFX28_2;
freq count;
model Y = X1;
run;
ODS HTML CLOSE;

```

【程序说明】 首先建立数据集， X_1 、 X_2 、 Y 和 $count$ 分别表示药物、医院、疗效和频数。接着调用 LOGISTIC 过程完成累积 logistic 回归模型的分析，其语句写法与二值变量的 logistic 回归相近。这里使用了两个 LOGISTIC 过程步，在第一个过程步中，class 语句用来指明 X_2 ，也就是医院为分类变量，使用该语句后，用户无需再去为医院这个多值名义变量设置哑变量。在第二个过程步中，由于对 X_2 进行检验无统计学意义，故将它去掉后重新拟合回归模型。

28.3.2 主要分析结果及结论

例 28-2 资料的主要输出结果如下：

| Analysis of Maximum Likelihood Estimates | | | | | | |
|--|-----|----------|----------------|-----------------|------------|--|
| Analysis of Maximum Likelihood Estimates | | | | | | |
| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq | |
| Intercept | 1 1 | -1.7939 | 0.2174 | 68.0711 | <.0001 | |
| Intercept | 2 1 | -0.0524 | 0.1786 | 0.0862 | 0.7691 | |
| Intercept | 3 1 | 1.4404 | 0.2112 | 46.5370 | <.0001 | |
| X1 | 1 | 0.6604 | 0.2413 | 7.4908 | 0.0062 | |
| X2 | 1 1 | -0.0530 | 0.2052 | 0.0667 | 0.7961 | |
| X2 | 2 1 | -0.0410 | 0.2064 | 0.0395 | 0.8425 | |
| X2 | 3 1 | 0.2918 | 0.2053 | 2.0204 | 0.1552 | |

第一个 LOGISTIC 过程步的输出结果中仅给出参数估计部分，其中变量 X_2 对应着 3 个参数估计值。LOGISTIC 过程在进行参数估计时，对于自变量为多值名义变量的情形，是以自变量的一个水平作为参照估计出多个参数， m 个水平的自变量可估计出 $m-1$ 个参数，默认状态下是以该自变量的最后一个水平作为参照。本例中医院有 4 个水平，所以有 3 个参数估计值，并且是以取值为 4 的医院作为参照来进行估计的。

The LOGISTIC Procedure

Model Information

| | |
|---------------------------|-------------|
| Data Set | WORK.PR28_2 |
| Response Variable | Y |
| Number of Response Levels | 4 |
| Frequency Variable | count |

| | |
|------------------------|------------------|
| Model | cumulative logit |
| Optimization Technique | Fisher's scoring |

Probabilities modeled are cumulated over the lower Ordered Values.

以上是第二个 LOGISTIC 过程步的第一部分输出结果，与二值变量 logistic 回归的不同之处在于这里使用的模型是累积 logit 模型，同时指出模型是建立在因变量排序较低的取值的发生概率上的，其他内容的说明可参见前例。

Score Test for the Proportional Odds Assumption

| | | |
|------------|----|------------|
| Chi-Square | DF | Pr > ChiSq |
| 1.1989 | 2 | 0.5491 |

第二部分输出的是平行线假设的结果，平行线假设也称为成比例发生比假设条件，在进行累积 logistic 回归时，需要满足该条件。这里 $\chi^2 = 1.1989$, $P = 0.5491 > 0.05$ ，说明该资料满足平行线假设条件。

Model Fit Statistics

| | | |
|-----------|----------------|--------------------------|
| Criterion | Intercept Only | Intercept and Covariates |
| AIC | 625.336 | 619.783 |
| SC | 635.689 | 633.587 |
| -2 Log L | 619.336 | 611.783 |

Testing Global Null Hypothesis: BETA = 0

| | | | |
|------------------|------------|----|------------|
| Test | Chi-Square | DF | Pr > ChiSq |
| Likelihood Ratio | 7.5537 | 1 | 0.0060 |
| Score | 7.4634 | 1 | 0.0063 |
| Wald | 7.4708 | 1 | 0.0063 |

第三部分输出模型拟合统计量和对整个模型进行检验的结果。在对整个模型进行的检验中，三种检验方法的结果都表明有统计学意义。

Analysis of Maximum Likelihood Estimates

Analysis of Maximum Likelihood Estimates

| | | | | | |
|-----------|-----|----------|----------------|-----------------|------------|
| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | 1 1 | -1.7767 | 0.2165 | 67.3251 | <.0001 |
| Intercept | 2 1 | -0.0499 | 0.1782 | 0.0784 | 0.7795 |
| Intercept | 3 1 | 1.4340 | 0.2108 | 46.2913 | <.0001 |
| X1 | 1 | 0.6589 | 0.2411 | 7.4708 | 0.0063 |

Odds Ratio Estimates

| | | |
|--------|----------------|----------------------------|
| Effect | Point Estimate | 95% Wald Confidence Limits |
| X1 | 1.933 | 1.205 3.100 |

第四部分是参数估计的结果，在累积 logistic 模型中，截距项有多个，其个数为因变量的水平数减 1，本例包含三个截距项。如果用 P_1 、 P_2 、 P_3 分别表示痊愈、显效、好转的概率，则回归方程为：

$$P_1 = \frac{e^{-1.7767+0.6589X_1}}{1 + e^{-1.7767+0.6589X_1}}$$

$$P_1 + P_2 = \frac{e^{-0.0499+0.6589X_1}}{1 + e^{-0.0499+0.6589X_1}}$$

$$P_1 + P_2 + P_3 = \frac{e^{1.4340+0.6589X_1}}{1 + e^{1.4340+0.6589X_1}}$$

专业结论：由以上结果可知，疗效与药物种类有关，而与医院无关。将 X_1 的取值代入上述方程，可发现：A 药的痊愈率和有效（即痊愈、显效和好转）率均高于 B 药，总体疗效优于 B 药。

28.4 多值名义变量的多重 logistic 回归分析

28.4.1 SAS 程序及说明

设对例 28-3 中资料进行多项 logit 模型分析的程序名为 SASTJFX28_3.SAS，程序如下：

```
data prg28_3;
  infile 'd:\sastjfx\fsxz.dat';
  input a b c y;
run;
ODS HTML STYLE = MINIMAL;
proc logistic data = prg28_3;

                                model y = a b c / link = glogit;
                                run;
                                proc logistic data = prg28_3;
                                model y = a b c / link = glogit noint;
                                run;
                                ODS HTML CLOSE;
```

【程序说明】 数据步是通过 infile 语句调用存放在 D 盘 sastjfx 文件夹内的外部数据文件，名为 fsxz.dat。

本例使用 LOGISTIC 过程实现多值名义变量的多项 logit 回归模型分析。这里需要特别指出的是，在 SAS 较早的版本中，LOGISTIC 过程无法实现多项 logit 模型，只能由 CATMOD 过程完成这一分析，在 SAS 9.1 及以后的版本中，LOGISTIC 过程已增加了这一功能，并且两个过程的输出结果是一致的。具体来讲，就是通过 model 语句中的 link 选项来完成的，link = glogit 表示拟合多项 logit 模型，其默认值是 logit。在默认状态下，对于二值变量 SAS 系统拟合一般的 logistic 模型，对于多值变量 SAS 系统拟合累积 logit 模型。这里使用了两个 LOGISTIC 过程步是因为在最初拟合的模型中截距项无统计学意义，所以在第二个 LOGISTIC 过程步中用 noint 选项去掉了截距项。另外，对于年龄这个多值有序变量，仍然作为连续变量处理。程序中其他语句的写法与前面的例子相同，在此不赘述。

28.4.2 主要分析结果及结论

由于两个 LOGISTIC 过程步的输出结果相近，这里只给出第二个 LOGISTIC 过程步的主要输出结果：

| The LOGISTIC Procedure | |
|---------------------------|-------------------|
| Model Information | |
| Data Set | WORK.PR28_3 |
| Response Variable | y |
| Number of Response Levels | 3 |
| Model | generalized logit |
| Optimization Technique | Fisher's scoring |

Logits modeled use y = 3 as the reference category.

以上是第二个 LOGISTIC 过程步的第一部分输出结果，这里使用的模型是广义 logit 模型，也就是多项 logit 模型。同时指出建模时是以 y = 3 为参照类估计方程，LOGISTIC 过程的默认方式是将结果变量的最后一个类别作为参照类，这里 3 就是变量 y 取值顺序中的最后一位，这一点要注意与前面的二值 logistic 模型以及累积 logit 模型有所区别。如果想要改变参照类，比如以避孕方式中的 IUD 为参照类别，只需要进行一定的设置即可。其他内容的说明可参见前例。

| Model Fit Statistics | | |
|----------------------|--------------------|-----------------|
| Criterion | Without Covariates | With Covariates |
| AIC | 219.722 | 79.986 |
| SC | 219.722 | 95.617 |
| -2 Log L | 219.722 | 67.986 |

Testing Global Null Hypothesis: BETA = 0

| Test | Chi-Square | DF | Pr > ChiSq |
|------------------|------------|----|------------|
| Likelihood Ratio | 151.7361 | 6 | < .0001 |
| Score | 120.9651 | 6 | < .0001 |
| Wald | 38.7688 | 6 | < .0001 |

第二部分是模型拟合统计量和对整个模型进行检验的结果。由拟合统计量可以看出，包含自变量的模型要优于只包含截距项的模型；在对整个模型进行的检验中，三种检验方法的结果都表明有统计学意义。

| Type 3 Analysis of Effects | | | |
|----------------------------|----|-----------------|------------|
| Effect | DF | Wald Chi-Square | Pr > ChiSq |
| a | 2 | 19.5313 | < .0001 |
| b | 2 | 24.8134 | < .0001 |
| c | 2 | 17.3033 | 0.0002 |

第三部分结果是对各个自变量所产生的效应进行分析的结果。这部分显示了各个自变量对结果变量的“整体”作用，该检验的原假设为原因变量对所有 logit 中的任何一个都没有作用。可以看出，年龄、生育史和收入这三个自变量对于结果变量的影响都有统计学意义。

| Analysis of Maximum Likelihood Estimates | | | | | | |
|--|---|----|----------|----------------|-----------------|------------|
| Parameter | y | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| a | 1 | 1 | 2.3270 | 1.0137 | 5.2700 | 0.0217 |
| a | 2 | 1 | -1.7960 | 0.7846 | 5.2402 | 0.0221 |
| b | 1 | 1 | 7.5813 | 1.6942 | 20.0248 | < .0001 |
| b | 2 | 1 | 4.7251 | 1.1723 | 16.2462 | < .0001 |
| c | 1 | 1 | -5.8682 | 1.6732 | 12.3009 | 0.0005 |
| c | 2 | 1 | 0.6638 | 0.6121 | 1.1763 | 0.2781 |

| Odds Ratio Estimates | | | | |
|----------------------|---|----------------|----------------------------|--------|
| Effect | y | Point Estimate | 95% Wald Confidence Limits | |
| a | 1 | 10.248 | 1.405 | 74.725 |
| a | 2 | 0.166 | 0.036 | 0.772 |

| | | | | |
|---|---|----------|--------|----------|
| b | 1 | >999.999 | 70.865 | >999.999 |
| b | 2 | 112.743 | 11.330 | >999.999 |
| c | 1 | 0.003 | <0.001 | 0.075 |
| c | 2 | 1.942 | 0.585 | 6.446 |

第四部分是参数估计的结果,作为因变量的避孕方式有三个水平,需要建立两个 logit 模型,每个 logit 模型对应一组参数,所以共有两组参数估计值。用 $P(y=1)$ 、 $P(y=2)$ 和 $P(y=3)$ 分别表示使用 IUD、口服避孕药和避孕套的概率,模型的表达式可写为:

$$\ln \left[\frac{P(y=1)}{P(y=3)} \right] = 2.3270a + 7.5813b - 5.8682c$$

$$\ln \left[\frac{P(y=2)}{P(y=3)} \right] = -1.7960a + 4.7251b + 0.6638c$$

此社区人群使用三种避孕方式的概率方程分别为:

$$P(y=1) = \frac{e^{2.3270a+7.5813b-5.8682c}}{1 + e^{2.3270a+7.5813b-5.8682c} + e^{-1.7960a+4.7251b+0.6638c}}$$

$$P(y=2) = \frac{e^{-1.7960a+4.7251b+0.6638c}}{1 + e^{2.3270a+7.5813b-5.8682c} + e^{-1.7960a+4.7251b+0.6638c}}$$

$$P(y=3) = \frac{1}{1 + e^{2.3270a+7.5813b-5.8682c} + e^{-1.7960a+4.7251b+0.6638c}}$$

另外,就生育史这个因素来讲,绝大多数未生育者的避孕方式是使用避孕套,而已生育者大多数选择 IUD 这种方式,两组人数相差悬殊,所以生育史这个因素在第一个 logit 中的优势比为 >999.999,这一点特别做出说明。

专业结论:结合实际资料可以看到,避孕方式的选择与年龄、生育史和收入都有关系。将 a、b、c 各水平的值代入上述三个概率方程即可计算三种避孕方式使用的概率(也可在程序中使用 output 语句将预测概率值保存至 SAS 数据集中),可以发现:年龄较大者、已生育者和低收入者更倾向于选择 IUD 这种避孕方式。为节省篇幅,各种条件组合下使用三种避孕方式的概率未予列出。

28.5 本章小结

本章主要介绍了如何运用 SAS 软件实现各种不同类型的非条件 logistic 回归,从数据结构的分析、方法选择、SAS 程序的写法及结果解释等方面做了详细的阐述。二值响应变量、多值有序响应变量和多值名义响应变量的 logistic 回归均可由 SAS 软件中的 LOGISTIC 过程或 CATMOD 过程来完成,本章只介绍了 LOGISTIC 过程的实现步骤,感兴趣的读者可以尝试使用 CATMOD 过程完成相应的分析。

(毛宗福 崔丹 李长平 柳伟伟)

第 29 章 配对设计定性资料多重 logistic 回归分析

logistic 回归按照资料的设计类型可以分为非条件 logistic 回归和条件 logistic 回归。在上一章中介绍了非条件 logistic 回归，即非配对设计资料的 logistic 回归。本章将按照配对比例的不同分别讲述如何使用 SAS 软件实现条件 logistic 回归分析，即配对设计资料的 logistic 回归。

29.1 问题、数据及统计分析方法的选择

29.1.1 问题与数据

【例 29-1】 为研究出现小于胎龄儿(出生体重低于同孕龄出生新生儿出生体重第 10 百分位数的新生儿)的影响因素，对研究对象的生育史和身高、体重进行匹配，采用 1:1 配对设计，搜集 80 对病例和对照的资料。研究中探讨的影响因素较多，本例题中仅选取 3 个影响因素，它们分别是：孕妇年龄(X_1)，小于 25 岁取值为 1、大于等于 25 岁小于 30 岁取值为 2、大于等于 30 岁取值为 3；孕妇受教育程度(X_2)，小学及以下取值为 1、中学取值为 2、大学及以上取值为 3；家庭人均月收入(X_3)，小于等于 800 元取值为 1、大于 800 元小于等于 1500 元取值为 2、大于 1500 元取值为 3。有关资料见表 29-1。

表 29-1 分娩小于胎龄儿影响因素的 1:1 配对病例对照研究资料

| 配对 编号 | 病例组 (Y=1) | | | 配对 编号 | 对照组 (Y=0) | | |
|----------|-----------------|--------------------|----------------------|----------|-----------------|-----------------|----------------------|
| | 年龄 (X_1) | 受教育程度 (X_2) | 家庭人均月 收入(X_3) | | 年龄 (X_1) | 受教育程度 (X_2) | 家庭人均月 收入(X_3) |
| 1 | 3 | 2 | 2 | 1 | 2 | 2 | 3 |
| 2 | 2 | 2 | 1 | 2 | 3 | 3 | 2 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 80 | 2 | 2 | 3 | 80 | 2 | 2 | 2 |

【例 29-2】 为探索小于胎龄儿的原因，收集了 226 名产妇的资料。其中，47 例产妇分娩小于胎龄儿，179 例产妇分娩正常儿。按照产妇年龄(A)进行配对，分析初潮年龄(X_1)、产妇身高(X_2)、孕期是否补铁(X_3)、补钙(X_4)、补锌(X_5)与小于胎龄儿是否有关系。令小于胎龄儿为 1($Y=1$)、正常体重儿为 0($Y=0$)。具体资料见表 29-2。

表 29-2 新生儿出生体重影响因素的 m:n 配对病例对照研究资料

| NO | Y | A | X_1 | X_2 | X_3 | X_4 | X_5 | NO | Y | A | X_1 | X_2 | X_3 | X_4 | X_5 |
|-----|-----|-----|-------|-------|-------|-------|-------|-----|-----|-----|-------|-------|-------|-------|-------|
| 1 | 1 | 20 | 14 | 162 | 1 | 1 | 1 | 114 | 0 | 24 | 14 | 156 | 0 | 1 | 0 |
| 2 | 1 | 20 | 14 | 158 | 0 | 0 | 0 | 115 | 0 | 24 | 16 | 165 | 0 | 1 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 112 | 0 | 24 | 14 | 163 | 0 | 1 | 0 | 225 | 0 | 30 | 12 | 152 | 0 | 0 | 0 |
| 113 | 0 | 24 | 16 | 168 | 0 | 1 | 0 | 226 | 0 | 30 | 13 | 150 | 0 | 0 | 0 |

29.1.2 对数据结构的分析

对例 29-1 资料而言,该资料中响应变量是新生儿出生体重分类(Y),出生体重低于同孕龄出生的新生儿出生体重第 10 百分位数的为小于胎龄儿,判断为病例组($Y=1$);新生儿出生体重高于同孕龄出生的新生儿出生体重第 10 百分位数的为对照组($Y=0$)。自变量有 3 个,均为多值有序变量,分别是孕妇年龄(X_1)、受教育程度(X_2)和家庭人均月收入(X_3)。此外,表 29-1 中还包含一个变量,那就是配对编号,这是成组设计中所没有的。

对例 29-2 资料而言,本研究的对象为某医院的产妇,要考察的影响因素包括初潮年龄、产妇身高、孕期补铁史、孕期补钙史、孕期补锌史。表 29-2 中的结果变量为新生儿出生体重分类,属于二值变量;影响因素中初潮年龄和产妇身高是定量变量,孕期补铁史、补钙史、补锌史都是二值变量;产妇年龄为配对变量;变量 NO 则是观察对象的编号。

29.1.3 分析目的与统计分析方法的选择

对例 29-1 资料而言,此研究为 1:1 配对病例对照研究,每个匹配组中有一个病例和一个对照。研究的目的在于考察多个危险因素对发生小于胎龄儿的影响,结果变量属于二值变量,故可以采用条件 logistic 回归进行分析。需要注意的是由于其为配对设计,所以不宜采用一般的非条件 logistic 回归模型分析该资料。

对例 29-2 资料而言,本例属于 $m:n$ 配对病例对照研究,由于按照产妇年龄进行配对,每个匹配组内病例和对照的配对比例无法统一,并且每个匹配组总的观察对象个数也不一定相同。想要考察多个自变量对小于胎龄儿发生的影响,自变量中既包括定量变量,也包括定性变量,分析方法应该选择条件 logistic 回归。

29.1.4 条件 logistic 回归分析简介

配对设计能够改善两组研究对象的齐同性,提高研究效率。配对的因素一般是年龄、性别等重要的混杂因素。最常见的配对形式是每个匹配组中有一个病例和若干个对照,称为 1: m 配对设计;当然,不同匹配组中病例和对照的人数也可以是任意的,也就是说不同匹配组中病例数与对照数的比例可以不相等,称为 $m:n$ 配对设计。讨论此类问题时,关心的是在某一给定的条件下某事件发生的概率,这一概率称为条件概率,所以将此类 logistic 回归称为条件 logistic 回归,将非配对设计资料的 logistic 回归称为非条件 logistic 回归。

29.2 1:1 配对设计定性资料的多重 logistic 回归分析

29.2.1 SAS 程序及说明

设对例 29-1 资料拟合 logistic 回归模型的程序名为 SASTJFX29_1.SAS,程序如下:

```
data ;
infile 'd:\sastjfx\xytle.dat';
input id x1 -x3 y;
run;
ODS HTML STYLE = MINIMAL;

proc logistic descending;
model y = x1 -x3 / selection
      = stepwise;
strata id;
run;
ODS HTML CLOSE;
```

【程序说明】 首先建立数据集，使用 infile 语句读入存放在 D 盘 sastjfx 文件夹内的外部数据文件，文件名为 xytle.dat。id 代表配对编号， X_1 、 X_2 、 X_3 和 Y 分别表示 3 个自变量与因变量。在条件 logistic 回归中代表配对编号的变量非常重要，在过程步中要用到它。在较早版本的 SAS 软件中，LOGISTIC 过程无法直接实现条件 logistic 回归，拟合条件 logistic 模型常常需要借助于 PHREG 过程，而在 SAS 9.1 及以后的版本中，LOGISTIC 过程也可以实现这一功能。本例采用 LOGISTIC 过程进行分析。语句“proc logistic”表示调用 LOGISTIC 过程，该语句中的“descending”选项要求按照降序输出结果，即输出 Y 取值较大时 ($Y = 1$) 的参数估计值。MODEL 语句中的选项“selection = stepwise”表示使用逐步法进行自变量的筛选，当然，实际中应多采用几种筛选方法进行自变量的选择。“strata id;”指定分层变量为 id，正是通过该语句实现了条件 logistic 回归。

29.2.2 主要分析结果及解释

对例 29-1 资料进行分析的主要输出结果如下：

| The LOGISTIC Procedure | |
|-------------------------------|----------------------|
| Conditional Analysis | |
| Model Information | |
| Data Set | WORK. |
| Response Variable | y |
| Number of Response Levels | 2 |
| Number of Strata | 80 |
| Model | binary logit |
| Optimization Technique | Newton-Raphson ridge |
| Probability modeled is y = 1. | |

结果第一部分输出模型基本信息，其中数据集名称为，响应变量为 Y，它有两个水平，分层的层数为 80，也就是有 80 对数据，使用的模型是二值 logit 模型，参数估计时的优化方法是 Newton-Raphson ridge 法。最后一行文字说明该模型是以变量 Y 的 1 水平为基础，也就是以出现小于胎龄儿的概率为基础建模。

| Stepwise Selection Procedure | | | | |
|------------------------------|--------------------|------------|------------------|-----------|
| Strata Summary | | | | |
| Response Pattern | y | | Number of Strata | Frequency |
| | 1 | 0 | | |
| 1 | 1 | 1 | 80 | 160 |
| Step 0. No covariates. | | | | |
| Residual Chi-Square Test | | | | |
| Chi-Square | DF | Pr > ChiSq | | |
| 30.0600 | 3 | < .0001 | | |
| Step 1. Effect x2 entered: | | | | |
| Model Fit Statistics | | | | |
| Criterion | Without Covariates | | With Covariates | |
| AIC | 110.904 | | 77.875 | |

| | | | |
|--|------------|------------|------------|
| SC | 110.904 | 80.950 | |
| -2 Log L | 110.904 | 75.875 | |
| Testing Global Null Hypothesis: BETA = 0 | | | |
| Test | Chi-Square | DF | Pr > ChiSq |
| Likelihood Ratio | 35.0288 | 1 | < .0001 |
| Score | 26.8431 | 1 | < .0001 |
| Wald | 13.9916 | 1 | 0.0002 |
| Residual Chi-Square Test | | | |
| Chi-Square | DF | Pr > ChiSq | |
| 3.8090 | 2 | 0.1489 | |

No effects for the model in Step 1 are removed.

第二部分是对自变量进行筛选的过程。首先给出的是关于分层情况的总结,病例和对照在每个匹配组中各有一个,有 80 个匹配组,共计 160 例观测。自变量的筛选过程仅有两步。第 0 步时,模型中只包含截距项,此时输出了残差 χ^2 检验的结果,其中 $\chi^2 = 30.0600$, $P < 0.0001$,说明仅含有截距项的回归模型对资料的拟合效果不好,有必要引入对因变量有影响的自变量。第 1 步时, X_2 进入了方程,输出结果包括模型拟合统计量、对整个模型进行检验的结果以及残差 χ^2 检验的结果。模型拟合统计量包括 AIC、SC 和 -2LogL ,这三个统计量取值越小,说明模型拟合越好。结果表明在包含自变量的模型中,这三个统计量的取值都小于不包含自变量的模型。在对整个模型进行的检验中,三种检验方法的结果都表明有统计学意义。残差 χ^2 检验的结果为 $\chi^2 = 3.8090$, $P = 0.1489 > 0.05$,说明此时模型拟合给定资料的效果较好。

No (additional) effects met the 0.05 significance level for entry into the model.

Summary of Stepwise Selection

| Step | Effect | | DF | Number In | Score | Chi-Square | Wald Chi-Square | Pr > ChiSq |
|------|---------|---------|----|-----------|---------|------------|-----------------|------------|
| | Entered | Removed | | | | | | |
| 1 | x2 | | 1 | 1 | 26.8431 | | | <.0001 |

第三部分是对筛选过程的总结,可以看出只有 X_2 进入了模型。

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|-----------|----|----------|----------------|-----------------|------------|
| x2 | 1 | -2.6985 | 0.7214 | 13.9916 | 0.0002 |

Odds Ratio Estimates

| Effect | Point Estimate | 95% Wald Confidence Limits |
|--------|----------------|----------------------------|
| x2 | 0.067 | 0.016 0.277 |

第四部分是参数估计的结果, X_2 的回归系数的估计值为 -2.6985 , 对其进行检验的 $\chi^2 = 13.9916$, $P = 0.0002$ 。同时还得到 X_2 的优势比为 0.067 , 其 95% 可信区间为 $(0.016, 0.277)$ 。因为是以变量 y 的 1 水平(小于胎龄儿)为基础建立模型的,而 X_2 的优势比小于 1,说明 X_2 是一个保护因素。

专业结论: 结合实际资料,可以得出孕妇受教育程度与小于胎龄儿的发生有关,受教育程度越高,分娩小于胎龄儿的风险越低。另外,根据现有资料,还不能认为孕妇年龄、家庭人均月收入与小于胎龄儿的发生存在关联。

29.3 m:n 配对设计定性资料的多重 logistic 回归分析

29.3.1 SAS 程序及说明

设对例 29-2 资料拟合 logistic 回归模型的程序名为 SASTJFX29_2.SAS, 程序如下:

```
data ;
  infile 'd:\sastjfx\matchmn.dat';
  input no y a x1 -x5;
run;
ODS HTML STYLE = MINIMAL;

proc logistic descending;
  model y = x1 -x5 / selection
          = stepwise;
  strata a;
run;
ODS HTML CLOSE;
```

【程序说明】 首先建立数据集, NO 代表观察对象的编号, $X_1 - X_5$ 以及 y 分别表示 5 个自变量与因变量, A 代表配对变量年龄。本例中 m:n 配对设计的条件 logistic 回归仍然由 logistic 过程完成, 其语句写法与前例中 1:1 配对设计条件 logistic 回归的格式基本相同。语句“proc logistic”中的“descending”选项要求按照降序输出结果。MODEL 语句中的选项“selection = stepwise”表示使用逐步法进行自变量的筛选。“strata a;”指定分层变量, 也就是说, 配对变量为 a。

29.3.2 主要分析结果及解释

对例 29-2 资料进行分析的主要输出结果如下:

| The LOGISTIC Procedure | | | |
|-------------------------------|--|--|----------------------|
| Conditional Analysis | | | |
| Model Information | | | |
| Data Set | | | WORK.PRG29_2 |
| Response Variable | | | y |
| Number of Response Levels | | | 2 |
| Number of Strata | | | 11 |
| Model | | | binary logit |
| Optimization Technique | | | Newton-Raphson ridge |
| Probability modeled is y = 1. | | | |

结果第一部分输出模型基本信息, 其中数据集名称为 prg29_2, 响应变量为 Y, 它有两个水平, 分层的层数为 11, 也就是按年龄进行配对后, 有 11 个匹配组, 使用的模型是二值 logit 模型, 参数估计时的优化方法是 Newton - Raphson ridge 法。最后一行文字说明该模型是以 Y 的 1 水平为基础, 也就是以出现小于胎龄儿的概率为基础建模。

| Stepwise Selection Procedure | | | | |
|------------------------------|---|----|------------------|-----------|
| Strata Summary | | | | |
| Response Pattern | y | | Number of Strata | Frequency |
| | 1 | 0 | | |
| 1 | 2 | 7 | 1 | 9 |
| 2 | 4 | 10 | 1 | 14 |

| | | | | |
|----|---|----|---|----|
| 3 | 2 | 13 | 1 | 15 |
| 4 | 3 | 14 | 1 | 17 |
| 5 | 5 | 13 | 1 | 18 |
| 6 | 6 | 16 | 1 | 22 |
| 7 | 2 | 21 | 1 | 23 |
| 8 | 4 | 20 | 1 | 24 |
| 9 | 5 | 22 | 1 | 27 |
| 10 | 6 | 22 | 1 | 28 |
| 11 | 8 | 21 | 1 | 29 |

Step 0. No covariates.

Residual Chi-Square Test

| Chi-Square | DF | Pr > ChiSq |
|------------|----|------------|
| 19.1052 | 5 | 0.0018 |

Step 1. Effect x2 entered:

Model Fit Statistics

| Criterion | Without Covariates | With Covariates |
|-----------|--------------------|-----------------|
| AIC | 192.318 | 176.159 |
| SC | 192.318 | 179.579 |
| -2 Log L | 192.318 | 174.159 |

Testing Global Null Hypothesis: BETA = 0

| Test | Chi-Square | DF | Pr > ChiSq |
|------------------|------------|----|------------|
| Likelihood Ratio | 18.1595 | 1 | < .0001 |
| Score | 17.6031 | 1 | < .0001 |
| Wald | 15.9721 | 1 | < .0001 |

Residual Chi-Square Test

| Chi-Square | DF | Pr > ChiSq |
|------------|----|------------|
| 1.9158 | 4 | 0.7512 |

No effects for the model in Step 1 are removed.

第二部分是对自变量进行筛选的过程。

首先给出的是关于分层情况的总结,具体来说就是结果中的“Strata Summary”部分。可以看到,本例中有 11 种不同的配对形式,每种配对形式中病例和对照的比例都不相同。例如在第二种配对形式中,有 4 个病例和 10 个对照。本例中的数据总共有 11 个匹配组,同时,在每个匹配组内,总的观察例数也不相同。第一个和最后一个匹配组的总例数分别为 9 例和 29 例。

自变量的筛选过程也仅有两步。第 0 步时,模型中只包含截距项,此时输出了残差 χ^2 检验的结果,其中 $\chi^2 = 19.1052$, $P = 0.0018 < 0.05$,说明仅含有截距项的回归模型对资料的拟合效果不好,有必要引入对因变量有影响的自变量。第 1 步时, X_2 进入了方程,输出结果包括模型拟合统计量、对整个模型进行检验的结果以及残差 χ^2 检验的结果。模型拟合统计量包括 AIC、SC 和 -2LogL ,在包含自变量的模型中,这三个统计量的取值都小于不包含自变量的模型。在对整个模型进行的检验中,三种检验方法的结果都表明有统计学意义。残差 χ^2 检验的结果为 $\chi^2 = 1.9158$, $P = 0.7512 > 0.05$,说明此时模型拟合给定资料的效果较好。

No(additional) effects met the 0.05 significance level for entry into the model.

Summary of Stepwise Selection

| Step | Effect | | DF | Number In | Score Chi-Square | Wald Chi-Square | Pr > ChiSq |
|------|---------|---------|----|-----------|------------------|-----------------|------------|
| | Entered | Removed | | | | | |
| 1 | x2 | | 1 | 1 | 17.6031 | | < .0001 |

第三部分是对筛选过程的总结，5 个自变量中只有 X₂进入了模型。

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|-----------|----|----------|----------------|-----------------|------------|
| x2 | 1 | -0.1563 | 0.0391 | 15.9721 | < .0001 |

Odds Ratio Estimates

| Effect | Point Estimate | 95% Wald Confidence Limits |
|--------|----------------|----------------------------|
| x2 | 0.855 | 0.792 0.923 |

第四部分是参数估计的结果，X₂的回归系数的估计值为 -0.1563，对其进行检验的 $\chi^2 = 15.9721$ ， $P < 0.0001$ 。同时还得到 X₂的优势比为 0.855，其 95% 可信区间为(0.792, 0.923)。因为是以变量 y 的 1 水平(小于胎龄儿)为基础建立模型的，而 X₂的优势比小于 1，说明 X₂是一个保护因素。

专业结论：结合实际资料，可以得出产妇身高与小于胎龄儿的发生有关，随着身高的增加，分娩小于胎龄儿的风险会降低。另外，根据现有资料，还不能认为产妇初潮年龄、孕期补充相应营养素与小于胎龄儿的发生存在关联。

29.4 本章小结

本章主要介绍了如何运用 SAS 软件实现不同类型的条件 logistic 回归，涉及的配对形式包括 1:1 配对和 m:n 配对。不同配对形式的条件 logistic 回归在 SAS 中实现时，其程序写法和输出结果大体相同。较高版本的 SAS 软件中，LOGISTIC 过程和 PHREG 过程都可以完成条件 logistic 回归，本章只讲述了 LOGISTIC 过程的实现步骤，感兴趣的读者可以使用 PHREG 过程对例题中的资料进行分析。

此外，还需特别指出的是，每个例题的分析结果及结论，其正确性都取决于研究设计的正确性。若研究设计有漏洞，则分析结果及结论并不能反映问题的真实情况。

(毛宗福 崔丹 李长平 柳伟伟)

第 30 章 原因变量为定量变量的判别分析

根据明确分类的受试对象(或样品)的多个定量指标的取值建立一个或多个关系式(即判别函数式,通常其具有一定程度的出错概率),再根据某种或某些规则,基于已建立的判别函数式实现对归属尚不明确的那些新个体的分类或判别,这样一种研究方法被称为判别分析。很显然,判别分析中的结果变量为分类变量(二分类变量或多分类变量)。本章将重点介绍原因变量为定量变量的各种判别分析方法。

30.1 实 例

30.1.1 问题与数据

【例 30-1】 某医生为研究舒张压与血浆胆固醇对冠心病的影响情况,随机抽取并测定了某地从事某特殊工作的 50~59 岁女工冠心病病人 15 例和正常人 16 例的舒张压(DBP)与血浆胆固醇(CHOL),见表 30-1。试分析测定结果,并提供对未知个体属于冠心病患者还是正常人的判断。

表 30-1 冠心病病人(第 1 组)与正常人(第 2 组)舒张压与血浆胆固醇的测定结果

| 编号 | DBP | CHOL | 组别 | 编号 | DBP | CHOL | 组别 |
|-----|-------|------|-----|-----|-------|------|-----|
| 1 | 9.86 | 5.18 | 1 | 1 | 10.66 | 2.07 | 2 |
| 2 | 13.33 | 3.73 | 1 | 2 | 12.53 | 4.45 | 2 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| | | | | 16 | 9.33 | 3.63 | 2 |

【例 30-2】 某人收集到脾虚症患者和非脾虚症者各 18 例,测定了血浆白蛋白含量(g/L) X_1 、血红蛋白含量(g/L) X_2 与玫瑰花形成细胞率(%) X_3 三项定量指标的取值,见表 30-2。试根据这三个定量指标的取值寻找到一些已经存在的规律,以便对疑似患有脾虚症的就诊者做出辅助诊断。假定有两位疑似患有脾虚症的受检者,他们的血浆白蛋白含量(g/L) X_1 ,血红蛋白含量(g/L) X_2 与玫瑰花形成细胞率(%) X_3 三项定量指标的测定结果分别为:

A: $X_1 = 34.8$, $X_2 = 113.4$, $X_3 = 514.6$; B: $X_1 = 27.3$, $X_2 = 77.6$, $X_3 = 375.4$ 。

试判定他们分别属于两类受试者中的哪一类。

表 30-2 脾虚症患者与对照者三项定量指标的测定结果

| 编号 | X_1 | X_2 | X_3 | X_1 | X_2 | X_3 |
|-----|-------|-------|-------|-------|-------|-------|
| | 脾虚症患者 | | | 对照者 | | |
| 1 | 27.5 | 78.5 | 381.5 | 35.5 | 119.5 | 530.5 |
| 2 | 27.5 | 78.5 | 371.5 | 36.3 | 120.5 | 540.5 |
| ... | ... | ... | ... | ... | ... | ... |
| 18 | 31.8 | 90.5 | 452.5 | 34.5 | 116.5 | 523.5 |

【例 30-3】 有一种鸢尾属植物，其品种分三类：第 1 类为 Setosa(有刚毛的)、第 2 类为 Versicolor(杂色的)和第 3 类为 Virginica(暂未找到相应的中文解释)。某研究者收集到该种植物的三类样品各 50 株，对每株植物测定 4 项定量指标的取值。 X_1 : Sepal Length(萼片长度, mm); X_2 : Sepal Width(萼片宽度, mm); X_3 : Petal Length(花瓣长度, mm); X_4 : Petal Width(花瓣宽度, mm)。其测定结果如表 30-3 所示(详细数据见后面的 SAS 程序, 该数据称为 Fisher(1936) Iris Data), 其中 G 为分类变量, 从左至右依次为第 1 类、第 2 类和第 3 类。试根据这种植物的多项定量指标的取值建立一种规则, 用于对分类尚不明确的该类植物的归属情况进行分类。假定有两株该种植物的新样品, 它们的 X_1 (萼片长度, mm)、 X_2 (萼片宽度, mm)、 X_3 (花瓣长度, mm)、 X_4 (花瓣宽度, mm)四项定量指标的测定结果分别为:

A: $X_1 = 6.3$, $X_2 = 3.0$, $X_3 = 5.8$, $X_4 = 2.1$; B: $X_1 = 4.7$, $X_2 = 3.6$, $X_3 = 1.4$, $X_4 = 0.3$ 。
试判定它们分别属于三类植物中的哪一类。

表 30-3 三类鸢尾属植物 4 项定量指标的测定结果

| 编号 | X_1 | X_2 | X_3 | X_4 | G | X_1 | X_2 | X_3 | X_4 | G | X_1 | X_2 | X_3 | X_4 | G |
|-----|-------|-------|-------|-------|-----|-------|-------|-------|-------|-----|-------|-------|-------|-------|-----|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | 1 | 7.0 | 3.2 | 4.7 | 1.4 | 2 | 6.3 | 3.3 | 6.0 | 2.5 | 3 |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 | 1 | 6.4 | 3.2 | 4.5 | 1.5 | 2 | 5.8 | 2.7 | 5.1 | 1.9 | 3 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 50 | 5.0 | 3.3 | 1.4 | 0.2 | 1 | 5.7 | 2.8 | 4.1 | 1.3 | 2 | 5.9 | 3.0 | 5.1 | 1.8 | 3 |

30.1.2 对数据结构的分析

对例 30-1 中表 30-1 所呈现的数据结构从形式上看很像“单因素 2 水平设计二元定量资料”，若这样理解资料，则两类受试者属于“原因”，而舒张压与血浆胆固醇这两项指标都应称为“观测结果”。事实上，研究者关心的是，在测定了这些定量指标的数值之后，希望判定未知归属的个体属于哪一类，即应将前面的“原因”与“结果”地位互换！所以，应将表 30-1 资料视为“结果变量为二值变量且具有两个定量原因变量(即自变量)的数据结构”。

例 30-2 和例 30-3 与例 30-1 相似，不再具体分析。表 30-2 资料应视为“结果变量为二值变量且具有三个定量原因变量(即自变量)的数据结构”，表 30-3 资料应视为“结果变量为三值名义变量且具有四个定量自变量的数据结构”。

30.1.3 分析目的与统计分析方法的选择

对这种数据结构的资料，人们最易选用的统计分析方法为“一般的多重 logistic 回归分析”。通常情况下，此法不仅可以用于筛选有统计学意义的变量，也可以用于对新个体的归属进行预测，但在对新个体的归属进行判别时，人们很少使用此法。故对本章资料而言，为了实现对新个体分类之目的，常选用一般判别分析。

30.2 原因变量为定量变量的判别分析简介

在医学研究和疾病防治工作中，经常会遇到需要根据观测到的资料对所研究的对象进行分类的问题。例如，在医疗诊断中，根据就诊者多项检验指标的取值情况来判断此人是否为某病患者。

判别分析的任务是根据已掌握的一批分类明确的样品，建立较好的判别函数，使产生错判的事例最少，进而对给定的新样品进行判别，判断它们来自于哪一个总体。

对判别分析处理的资料而言,其结果变量均为定性资料,但原因变量则不然。根据原因变量的性质,可将判别分析分为原因变量为定量变量的判别分析和原因变量为定性资料的判别分析。本章主要介绍如何借助 SAS 软件实现原因变量为定量变量的判别分析,而原因变量为定性资料的判别分析将在第 31 章中予以介绍。

30.3 原因变量为定量变量的判别分析

30.3.1 SAS 程序

对于例 30-1 资料,可以运用下面的 SAS 程序实现判别分析。程序名为 SASTJFX30_1.SAS。

| | |
|--|---|
| <pre>data da; input nob dbp chol g @@ ; cards; 1 9.86 5.18 1 1 10.66 2.07 2 2 13.33 3.73 1 2 12.53 4.45 2 3 14.66 3.89 1 3 13.33 3.06 2 4 9.33 7.10 1 4 9.33 3.94 2 5 12.80 5.49 1 5 10.66 4.45 2 6 10.66 4.09 1 6 10.66 4.92 2 7 10.66 4.45 1 7 9.33 3.68 2 8 13.33 3.63 1 8 10.66 2.77 2 9 13.33 5.96 1 9 10.66 3.21 2 10 13.33 5.70 1 10 10.66 5.02 2 11 12.00 6.19 1 11 10.40 3.94 2 12 14.66 4.01 1 12 9.33 4.92 2 13 13.33 4.01 1 13 10.66 2.69 2 14 12.80 3.63 1 14 10.66 2.43 2 15 13.33 5.96 1 15 11.20 3.42 2 16 9.33 3.63 2 ; run;</pre> | <pre>ODS HTML; PROC DISCRIM data = da METHOD = NORMAL POOL = TEST MANOVA CROSSTESTERR; CLASS g; VAR dbp chol; quit; PROC DISCRIM data = da METHOD = NPAR K = 8 LISTERR CROSSTESTERR; CLASS g; VAR dbp chol; quit; ODS HTML CLOSE;</pre> |
|--|---|

【程序说明】 数据步很简单,不做说明。第一个过程步中调用的是一般判别分析过程,其过程名为 DISCRIM。该过程对定量的自变量不进行变量筛选,将定量自变量全部纳入计算。该过程中能采用参数法判别分析和非参数法判别分析,其关键词是“METHOD =”。若在等号后填写“NORMAL”,表明希望采用参数法进行判别分析,其前提条件是各类定量资料服从多元正态分布;若在等号后填写“NPAR”,表明希望采用非参数法进行判别分析。

在“METHOD = NORMAL”之后,“POOL =”的等号之后有三种写法,它们分别是 NO(对类间协方差矩阵不进行假设检验,直接采用各类的协方差矩阵计算类间距离,最后将会给出二次型判别函数式)、YES(求所有类间合并的协方差矩阵,并据此计算类间广义平方距离,最后将给出线性判别函数式)和 TEST(要求对类间协方差矩阵是否相等进行假设检验,采用的是似然比检验的 Bartlett's 校正法。若相等,则对各类协方差矩阵进行合并计算,相当于 YES 的结果;若不相等,则相当于 NO 的结果)。

MANOVA 要求系统对两类间定量资料进行多元方差分析(本例中就是进行单因素 2 水平设计定量资料二元方差分析);CROSSTESTERR 要求系统列出交叉验证的判别结果中判断错误的那些观测。

第二个过程步中仍然调用了一般判别分析过程，但有两个选项做了改变，即“METHOD = NPAR K=8”，意思是选用非参数判别分析中的 K 最近邻法，并取 K=8。K 取什么值合适，需要尝试，通常从 1 开始尝试，直到获得最小的交叉验证误判率为止。

对于例 30-2 资料，可以运用下面的 SAS 程序实现判别分析。程序名为 SASTJFX30_2.SAS。

| | |
|---|---|
| <pre>DATA da; DO q=1 TO 2; INPUT x1 -x3 @@ ; OUTPUT; END; CARDS; 27.5 78.5 381.5 35.5 119.5 530.5 27.5 78.5 371.5 36.3 120.5 540.5 27.5 80.5 381.5 38.5 127.5 541.5 29.5 80.5 381.5 37.4 126.5 541.2 28.5 79.5 401.5 36.3 120.5 540.8 28.5 80.5 405.2 35.4 118.5 530.5 30.5 88.5 412.5 34.5 110.5 522.5 31.5 87.5 442.5 36.5 123.5 530.5 30.5 87.5 422.5 34.2 109.2 503.5 30.1 88.5 415.2 34.6 108.5 513.2 31.2 91.5 433.5 33.5 105.3 510.8 30.4 80.5 405.5 34.5 115.2 520.5 30.5 91.5 415.5 35.5 120.5 530.5 30.6 87.5 405.4 35.0 118.2 525.2 30.9 83.5 420.5 35.0 118.0 530.3 31.1 88.5 430.5 35.3 118.2 528.0 31.6 91.5 445.2 35.2 117.5 524.2 31.8 90.5 452.5 34.5 116.5 523.5 ; RUN;</pre> | <pre>data newdata; input X1 -X3; cards; 34.8 113.4 514.6 27.3 77.6 375.4 ; RUN; ODS HTML; PROC DISCRIM data=da testdata=newdata METHOD=NORMAL POOL=TEST MANOVA CROSSLISTERR OUTSTAT=aaa TESTOUTD=bbb; CLASS g; VAR x1 -x3; quit; PROC PRINT data=aaa; where _TYPE_='QUAD'; RUN; PROC PRINT data=bbb; RUN; ODS HTML CLOSE;</pre> |
|---|---|

【程序说明】 第一个数据步录入表 30-2 中的数据，保存为数据集 da。第二个数据步录入待判别类型的新数据，保存为数据集 newdata。

第一个过程步 (DISCRIM 过程) 中与程序 SASTJFX30_1 相似的选项的含义，不再赘述。“testdata=”选项等号后填写包含新个体的 SAS 数据集名，“testoutd=”选项等号后填写由对新个体判别分类的计算结果所形成的 SAS 数据集，“OUTSTAT=”选项用来产生包含各种计算结果的数据集。之后的两个 print 过程分别将判别函数式中的系数和对新数据集的判别结果输出到 output 窗口。由 DISCRIM 过程步的输出结果(见 30.3.2 节)可得到，本资料的判别函数式是二次型的，而数据集 aaa 中包含内容众多，我们仅希望输出二次型判别函数式中的系数，故在此过程步中加上一个 WHERE 语句。

对于例 30-3 资料，可以运用下面的 SAS 程序实现判别分析。程序名为 SASTJFX30_3.SAS。

| | |
|---|--|
| <pre>proc format; value specname 1 = 'Setosa' 2 = 'Versicolor' 3 = 'Virginica'; run; data iris;</pre> | <pre>data newdata; input SepalLength SepalWidth PetalLength PetalWidth; cards; 6.3 3.0 5.8 2.1 4.7 3.6 1.4 0.3 ;</pre> |
|---|--|

```

title 'Fisher(1936) Iris Data';
infile
'd:\sastjfx\P657_data_on_irises.
dat';
input SepalLength SepalWidth
PetalLength PetalWidth Species;
format Species specname.;
label
SepalLength = 'Sepal Length in mm.'
SepalWidth = 'Sepal Width in mm.'
PetalLength = 'Petal Length in mm.'
PetalWidth = 'Petal Width in mm.';
symbol =put (Species,specname10.);
run;

run;
ODS HTML;
PROC DISCRIM data =iris
testdata =newdata METHOD =NPAR
K =6 MANOVA LISTERR CROSSLISTERR
testoutd =bbb;
CLASS Species;
VAR SepalLength SepalWidth
PetalLength PetalWidth;
RUN;
PROC PRINT data =bbb;
run;
ODS HTML CLOSE;

```

【程序说明】 第一个过程步用来定义分类变量的格式，本例中，分类变量为植物种类 (species)，它有三个水平，分别为 Setosa、Versicolor 和 Virginica，在数据文件中分别用 1、2 和 3 来表示，而希望在输出结果中用真实的分类名称表示，故需要用此过程来定义输出格式。

将表 30-3 所代表的完整的原始数据以文本格式存在 D 盘 sastjfx 文件夹内，文件名为“P657_data_on_irises.dat”。由数据步中的“infile”语句将其打开。

过程步中采用了非参数判别分析 (METHOD = NPAR)，具体方法叫 K 最近邻法 (K = 6)，其中“K = ”后可尝试填 1, 2, 3, … 正整数。本例经尝试后发现 K 取 6 时，交叉验证的误判率最低。

30.3.2 主要分析结果及解释

借助 SAS 软件对例 30-1 资料进行判别分析的结果如下。

| The DISCRIM Procedure | | |
|--------------------------------------|---------------------------|--|
| Within Covariance Matrix Information | | |
| g | Covariance Matrix Rank | Natural Log of the Determinant of the Covariance Matrix |
| 1 | 2 | 1.04756 |
| 2 | 2 | 0.00379 |
| Pooled | 2 | 0.60475 |

以上是关于两组内部及合并的协方差矩阵的行列式值的自然对数值。

| Test of Homogeneity of Within Covariance Matrices | | | |
|---|----|------------|--|
| Chi-Square | DF | Pr > ChiSq | |
| 2.604266 | 3 | 0.4567 | |

Since the Chi-Square value is not significant at the 0.1 level, a pooled covariance matrix will be used in the discriminant function.

Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

以上是关于两组协方差矩阵是否满足齐性要求的假设检验结果， $\chi^2 = 2.604$ ， $df = 3$ ， $P = 0.4567$ 。说明两组协方差矩阵满足齐性要求，可用合并的协方差矩阵求判别函数式中的系数。

| Generalized Squared Distance to g | | |
|-----------------------------------|---|---|
| From g | 1 | 2 |

| | | |
|---|---------|---------|
| 1 | 0 | 4.64167 |
| 2 | 4.64167 | 0 |

以上给出了两组间广义平方距离为 4.642。

Multivariate Statistics and Exact F Statistics

S = 1 M = 0 N = 13

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|------------------------|------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.44659674 | 17.35 | 2 | 28 | <.0001 |
| Pillai's Trace | 0.55340326 | 17.35 | 2 | 28 | <.0001 |
| Hotelling-Lawley Trace | 1.23915652 | 17.35 | 2 | 28 | <.0001 |
| Roy's Greatest Root | 1.23915652 | 17.35 | 2 | 28 | <.0001 |

以上用了四种单因素 2 水平设计定量资料二元方差分析,目的是检验两组受试者在两个定量指标组成的平均向量之间的差别是否具有统计学意义。通常看第一种方法给出的分析结果即可。本例 Wilks' $\lambda = 0.446597$, 对应的 $F = 17.35$, $df_{\text{分子}} = 2$, $df_{\text{分母}} = 28$, $P < 0.0001$, 说明用给定的两个定量指标能较好地地区分冠心病人与正常人。

Linear Discriminant Function for g

| Variable | 1 | 2 |
|----------|-----------|-----------|
| Constant | -72.49746 | -49.25511 |
| dbp | 8.41761 | 7.04324 |
| chol | 8.18104 | 6.45716 |

这里给出了两个判别函数中的系数估计值,代表第一类与第二类受试者的判别函数分别为:

$$y_1 = -72.49746 + 8.41761\text{dbp} + 8.18104\text{chol}$$

$$y_2 = -49.25511 + 7.04324\text{dbp} + 6.45716\text{chol}$$

Classification Summary for Calibration Data: WORK.DA

Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent

| Classified into g | | | |
|-------------------|-------|-------|--------|
| From g | 1 | 2 | Total |
| 1 | 12 | 3 | 15 |
| | 80.00 | 20.00 | 100.00 |
| 2 | 3 | 13 | 16 |
| | 18.75 | 81.25 | 100.00 |
| Total | 15 | 16 | 31 |
| | 48.39 | 51.61 | 100.00 |
| Priors | 0.5 | 0.5 | |

以上是利用已求得的线性判别函数进行回代判别的结果:原先为第一类仍判为第一类的有 12 例;原先为第二类仍判为第二类的有 13 例。

Error Count Estimates for g

| | 1 | 2 | Total |
|--------|--------|--------|--------|
| Rate | 0.2000 | 0.1875 | 0.1938 |
| Priors | 0.5000 | 0.5000 | |

以上是误判率：两类共有 31 例，出现误判的有 6 例，两类平均误判率为 19.38%。

说明：这里的平均误判率是将两类各自误判率相加后除以 2 所得，此计算方法欠妥，应该用总误判数除以总例数，即 $(6/31) \times 100\% = 19.35\%$ 。

| Classification Results for Calibration Data: WORK. DA | | | | | |
|---|--------|-------------------|---|--------|--------|
| Cross-validation Results using Linear Discriminant Function | | | | | |
| Posterior Probability of Membership | | | | | |
| in g | | | | | |
| Obs | From g | Classified into g | 1 | 2 | |
| 1 | 1 | 2 | * | 0.2111 | 0.7889 |
| 4 | 2 | 1 | * | 0.8908 | 0.1092 |
| 6 | 2 | 1 | * | 0.7122 | 0.2878 |
| 11 | 1 | 2 | * | 0.1228 | 0.8772 |
| 12 | 2 | 1 | * | 0.5152 | 0.4848 |
| 13 | 1 | 2 | * | 0.2318 | 0.7682 |
| 20 | 2 | 1 | * | 0.5671 | 0.4329 |
| * Misclassified observation | | | | | |

以上是基于交叉验证法(所谓交叉验证，也被称为刀切法，即每次丢下一个受试者的数据后创建判别函数式，再依据所求得的判别函数式对未参与计算的这个受试者进行判别，数据集中的每位受试者都要被丢下一次。通常情况下，交叉验证法所得到的判别准确率会有所下降)给出的误判受试者的列表，每一行对应 1 个被误判的受试者，一共有 7 个受试者被误判了。其中第一类有 3 例被误判(“From g”下的取值为“1”)，第二类有 4 例被误判(“From g”下的取值为“2”)，判别的依据是依据线性判别函数计算出属于第一类和第二类的概率(上面结果中最后两列小数)，判为概率大的那一类。

| Classification Summary for Calibration Data: WORK. DA | | | |
|---|--------|--------|--------|
| Cross-validation Summary using Linear Discriminant Function | | | |
| Number of Observations and Percent | | | |
| Classified into g | | | |
| From g | 1 | 2 | Total |
| 1 | 12 | 3 | 15 |
| | 80.00 | 20.00 | 100.00 |
| 2 | 4 | 12 | 16 |
| | 25.00 | 75.00 | 100.00 |
| Total | 16 | 15 | 31 |
| | 51.61 | 48.39 | 100.00 |
| Priors | 0.5 | 0.5 | |
| Error Count Estimates for g | | | |
| Error Count Estimates for g | | | |
| | 1 | 2 | Total |
| Rate | 0.2000 | 0.2500 | 0.2250 |
| Priors | 0.5000 | 0.5000 | |

以上给出了基于交叉验证法判别的结果, 其中第一类误判了 3 例, 第二类误判了 4 例, 平均误判率为 22.50% (这个平均误判率也是将两类各自的误判率相加后除以 2 所得, 此计算方法欠妥, 应该用总误判数除以总例数, 即 $(7/31) \times 100\% = 22.58\%$)。

第二个过程步(即 K 最近邻法进行非参数判别分析)产生的输出结果如下。

| Classification Results for Calibration Data: WORK. DA | | | | | |
|---|--------|------------|--------|--------|--------|
| Resubstitution Results using 8 Nearest Neighbors | | | | | |
| Posterior Probability of Membership | | | | | |
| in g | | | | | |
| Obs | From g | Classified | into g | 1 | 2 |
| 1 | 1 | 2 | * | 0.3902 | 0.6098 |
| 4 | 2 | 1 | * | 0.6400 | 0.3600 |
| 6 | 2 | 1 | * | 0.6400 | 0.3600 |
| 11 | 1 | 2 | * | 0.2623 | 0.7377 |
| 13 | 1 | 2 | * | 0.3902 | 0.6098 |

* Misclassified observation

以上是基于回代判别而产生误判的 5 个受试者的编号, 第一类误判 3 例, 第二类误判 2 例。

| Classification Summary for Calibration Data: WORK. DA | | | |
|---|-------|-------|--------|
| Resubstitution Summary using 8 Nearest Neighbors | | | |
| Number of Observations and Percent | | | |
| Classified into g | | | |
| From g | 1 | 2 | Total |
| 1 | 12 | 3 | 15 |
| | 80.00 | 20.00 | 100.00 |
| 2 | 2 | 14 | 16 |
| | 12.50 | 87.50 | 100.00 |
| Total | 14 | 17 | 31 |
| | 45.16 | 54.84 | 100.00 |
| Priors | 0.5 | 0.5 | |

| Error Count Estimates for g | | | |
|-----------------------------|--------|--------|--------|
| | 1 | 2 | Total |
| Rate | 0.2000 | 0.1250 | 0.1625 |
| Priors | 0.5000 | 0.5000 | |

以上是基于回代判别所产生的平均误判率为 16.25% (此平均误判率仍是将两类各自的误判率相加后除以 2 所得, 此计算方法欠妥, 应该用总误判数除以总例数, 即 $(5/31) \times 100\% = 16.13\%$)。

| Classification Results for Calibration Data: WORK. DA | | | | | |
|---|--------|------------|--------|--------|--------|
| Cross-validation Results using 8 Nearest Neighbors | | | | | |
| Posterior Probability of Membership | | | | | |
| in g | | | | | |
| Obs | From g | Classified | into g | 1 | 2 |
| 1 | 1 | 2 | * | 0.2759 | 0.7241 |
| 4 | 2 | 1 | * | 0.6250 | 0.3750 |

| | | | | | |
|----|---|---|---|--------|--------|
| 6 | 2 | 1 | * | 0.7500 | 0.2500 |
| 11 | 1 | 2 | * | 0.1404 | 0.8596 |
| 13 | 1 | 2 | * | 0.2759 | 0.7241 |

* Misclassified observation

以上是基于交叉验证法判别而产生误判的 5 个受试者的编号, 第一类误判 3 例, 第二类误判 2 例。

Classification Summary for Calibration Data: WORK.DA

Cross-validation Summary using 8 Nearest Neighbors

Number of Observations and Percent

Classified into g

| From g | 1 | 2 | Total |
|--------|-------|-------|--------|
| 1 | 12 | 3 | 15 |
| | 80.00 | 20.00 | 100.00 |
| 2 | 2 | 14 | 16 |
| | 12.50 | 87.50 | 100.00 |
| Total | 14 | 17 | 31 |
| | 45.16 | 54.84 | 100.00 |
| Priors | 0.5 | 0.5 | |

Error Count Estimates for g

| | 1 | 2 | Total |
|--------|--------|--------|--------|
| Rate | 0.2000 | 0.1250 | 0.1625 |
| Priors | 0.5000 | 0.5000 | |

以上是基于交叉验证法判别所产生的平均误判率为 16.25% (此平均误判率仍是将两类各自的误判率相加后除以 2 所得, 此计算方法欠妥, 应该用总误判数除以总例数, 即 $(5/31) \times 100\% = 16.13\%$)。

结论: 本例分别用参数法和非参数法进行了判别分析, 相对来说, 非参数法的误判率稍低一点。很可能是本资料不符合参数判别分析的前提条件(应服从二元正态分布)。

借助 SAS 软件对例 30-2 资料进行判别分析的结果如下。

The DISCRIM Procedure

| | | | |
|--------------|----|--------------------|----|
| Observations | 36 | DF Total | 35 |
| Variables | 3 | DF Within Classes | 34 |
| Classes | 2 | DF Between Classes | 1 |

Class Level Information

| g | Variable Name | Frequency | Weight | Proportion | Prior Probability |
|---|---------------|-----------|---------|------------|-------------------|
| 1 | _1 | 18 | 18.0000 | 0.500000 | 0.500000 |
| 2 | _2 | 18 | 18.0000 | 0.500000 | 0.500000 |

以上是关于资料的基本信息, 即总共 36 个观测(即样品数), 变量个数为 3, 分类个数为 2, 总自由度为 35, 类内自由度为 34, 类间自由度为 1。两类的观测个数均为 18, 各类占总样品数的比例均为 0.5, 各类样品出现的事前概率均为 0.5。因为在程序中没有事先给定各类样品出现的概率值的大小, SAS 软件包中规定的缺省值为 0.5。

若根据专业知识得知,第1类与第2类样品出现的概率分别为0.6与0.4,则可以将过程步修改(即增加了倒数第2句)如下:

```
PROC DISCRIM data = da METHOD = NORMAL POOL = TEST
    MANOVA SHORT CROSSTESTERR;
    CLASS g;
    VAR x1 - x3;
    PRIORS '1' = 0.6 '2' = 0.4;
quit;
```

(注意:下面的输出结果仍基于前面的SAS程序!)

Test of Homogeneity of Within Covariance Matrices

| Chi-Square | DF | Pr > ChiSq |
|------------|----|------------|
| 25.305199 | 6 | 0.0003 |

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.

Reference: Morrison, D. F. (1976) Multivariate Statistical Methods p252.

以上给出了两类数据的协方差矩阵是否相等的假设检验结果,采用 χ^2 检验,得 $\chi^2 = 25.305$, $df = 6$, $P = 0.0003 < 0.05$, 拒绝 H_0 (两类数据的协方差矩阵相等), 接受 H_1 (两类数据的协方差矩阵不相等)。

Multivariate Statistics and Exact F Statistics

S = 1 M = 0.5 N = 15

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|------------------------|-------------|---------|--------|--------|--------|
| Wilks' Lambda | 0.05765212 | 174.35 | 3 | 32 | <.0001 |
| Pillai's Trace | 0.94234788 | 174.35 | 3 | 32 | <.0001 |
| Hotelling-Lawley Trace | 16.34541496 | 174.35 | 3 | 32 | <.0001 |
| Roy's Greatest Root | 16.34541496 | 174.35 | 3 | 32 | <.0001 |

以上是将原始数据的结构视为“单因素2水平设计三元定量资料”并采用单因素2水平设计三元定量资料方差分析处理资料,目的是考察两类受试者同时在这三个定量指标的平均值之间的差别是否具有统计学意义。SAS系统中采用了四种实现多元方差分析的方法,通常看第一种计算结果即可。本例Wilks' $\lambda = 0.05765212$,其对应的F检验统计量为 $F = 174.35$,分子与分母的自由度分别为 $df_{\text{分子}} = 3$ 、 $df_{\text{分母}} = 32$,对应的P值为 $P < 0.0001$ 。说明用这三个定量指标能比较好地区分任何一个属于本研究问题中的受试者究竟属于脾虚症患者还是对照者。不过,由于前述结果已经表明两类数据的协方差矩阵不等,资料不满足参数检验的前提条件,故此处的方差分析结果仅起参考作用。

Classification Summary for Calibration Data: WORK.DA

Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent

| | | Classified into g | | |
|--------|--|-------------------|--------|--------|
| From g | | 1 | 2 | Total |
| 1 | | 18 | 0 | 18 |
| | | 100.00 | 0.00 | 100.00 |
| 2 | | 0 | 18 | 18 |
| | | 0.00 | 100.00 | 100.00 |

| | | | |
|--------|-------|-------|--------|
| | 18 | 18 | 36 |
| Total | 50.00 | 50.00 | 100.00 |
| Priors | 0.5 | 0.5 | |

以上是将原先的每位受试者在三项定量指标上的取值代入所求得的二次型判别函数式(这种判别函数式的系数项比较多,需要通过产生输出数据文件才能得到,后面详述)进行计算,并判断其属于哪一类所得到的“回代判别”结果。左边“From g”代表原先的分类,表头上“into g”代表计算以后的分类。不难看出,两类受试者全部都被正确地分了类,分类正确率为 100%。

| | | | |
|-----------------------------|--------|--------|--------|
| Error Count Estimates for g | | | |
| | 1 | 2 | Total |
| Rate | 0.0000 | 0.0000 | 0.0000 |
| Priors | 0.5000 | 0.5000 | |

这一部分给出了各类的误判率均为 0; 各类事前概率均为 0.5(SAS 系统中的缺省值,因为在第一个 SAS 程序中并没有人为指定其取值)。

Classification Summary for Calibration Data: WORK.DA
Cross-validation Summary using Quadratic Discriminant Function

| | | | |
|------------------------------------|--------|--------|--------|
| Number of Observations and Percent | | | |
| Classified into g | | | |
| From g | 1 | 2 | Total |
| 1 | 18 | 0 | 18 |
| | 100.00 | 0.00 | 100.00 |
| 2 | 0 | 18 | 18 |
| | 0.00 | 100.00 | 100.00 |
| Total | 18 | 18 | 36 |
| | 50.00 | 50.00 | 100.00 |
| Priors | 0.5 | 0.5 | |

这是利用所求得的二次型判别函数式进行交叉验证的计算结果,左边“From g”代表原先的分类,表头上“into g”代表计算以后的分类。不难看出,两类受试者全部都被正确地分了类,分类正确率为 100%。

| | | | |
|-----------------------------|--------|--------|--------|
| Error Count Estimates for g | | | |
| | 1 | 2 | Total |
| Rate | 0.0000 | 0.0000 | 0.0000 |
| Priors | 0.5000 | 0.5000 | |

这一部分给出了基于交叉验证的各类的误判率均为 0; 各类事前概率均为 0.5。

| Obs | g | _TYPE_ | _NAME_ | x1 | x2 | x3 |
|-----|---|--------|----------|---------|---------|---------|
| 67 | 1 | QUAD | x1 | -1.40 | 0.13 | 0.06 |
| 68 | 1 | QUAD | x2 | 0.13 | -0.08 | 0.01 |
| 69 | 1 | QUAD | x3 | 0.06 | 0.01 | -0.01 |
| 70 | 1 | QUAD | _LINEAR_ | 15.83 | 0.39 | -0.20 |
| 71 | 1 | QUAD | _CONST_ | -216.38 | -216.38 | -216.38 |
| 72 | 2 | QUAD | x1 | -1.89 | 0.28 | 0.04 |
| 73 | 2 | QUAD | x2 | 0.28 | -0.11 | 0.03 |

| | | | | | | |
|----|---|------|----------|----------|----------|----------|
| 74 | 2 | QUAD | x3 | 0.04 | 0.03 | -0.02 |
| 75 | 2 | QUAD | _LINEAR_ | 24.16 | -23.54 | 14.04 |
| 76 | 2 | QUAD | _CONST_ | -2749.39 | -2749.39 | -2749.39 |

这里给出了二次型判别函数式中的系数。依据此结果，可以写出如下两个二次型判别函数式：

$$\begin{aligned}y_1 &= -216.38 + 15.83X_1 + 0.39X_2 - 0.20X_3 - 1.40X_1^2 \\&\quad - 0.08X_2^2 - 0.01X_3^2 + 0.13X_1X_2 + 0.06X_1X_3 + 0.01X_2X_3 \\y_2 &= -2749.39 + 24.16X_1 - 23.54X_2 + 14.04X_3 - 1.89X_1^2 \\&\quad - 0.11X_2^2 - 0.02X_3^2 + 0.28X_1X_2 + 0.04X_1X_3 + 0.03X_2X_3\end{aligned}$$

| Obs | X1 | X2 | X3 | _1 | _2 |
|-----|------|-------|-------|------------|------------|
| 1 | 34.8 | 113.4 | 514.6 | 7.9595E-13 | .001390638 |
| 2 | 27.3 | 77.6 | 375.4 | .000293888 | 3.8977E-86 |

这是依据所求得的二次型判别函数式计算给定的两个新个体应判为哪一类的结果，将个体 A 判为第一类所对应的数值很小，判为第二类所对应的数值相对较大，故应将其判为第二类，即为对照组；同理，应判个体 B 为第一类，即为脾虚症患者。

借助 SAS 软件对例 30-3 资料进行判别分析的结果如下。

| Fisher(1936) Iris Data | | | | | |
|--|-------------|---------|--------|--------|--------|
| Multivariate Statistics and F Approximations | | | | | |
| S = 2 M = 0.5 N = 71 | | | | | |
| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
| Wilks' Lambda | 0.02343863 | 199.15 | 8 | 288 | <.0001 |
| Pillai's Trace | 1.19189883 | 53.47 | 8 | 290 | <.0001 |
| Hotelling-Lawley Trace | 32.47732024 | 582.20 | 8 | 203.4 | <.0001 |
| Roy's Greatest Root | 32.19192920 | 1166.96 | 4 | 145 | <.0001 |

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

以上是单因素三水平设计定量资料四元方差分析结果，本例 Wilks' $\lambda = 0.02343863$ ，其对应的 F 检验统计量为 $F = 199.15$ ，分子与分母的自由度分别为 $df_{\text{分子}} = 8$ 、 $df_{\text{分母}} = 288$ ，对应的 P 值为 $P < 0.0001$ 。说明用这四个定量指标能比较好地区分任何一个属于本研究问题中的植物究竟属于三类中的哪一类。

| Classification Results for Calibration Data: WORK. IRIS | | | | | | |
|---|--------------|-------------------------|---|--------|------------|-----------|
| Resubstitution Results using 6 Nearest Neighbors | | | | | | |
| Posterior Probability of Membership in Species | | | | | | |
| Obs | From Species | Classified into Species | | Setosa | Versicolor | Virginica |
| 71 | Versicolor | Other | T | 0.0000 | 0.5000 | 0.5000 |
| 73 | Versicolor | Other | T | 0.0000 | 0.5000 | 0.5000 |
| 84 | Versicolor | Virginica | * | 0.0000 | 0.1667 | 0.8333 |
| 134 | Virginica | Versicolor | * | 0.0000 | 0.6667 | 0.3333 |

* Misclassified observation; T Tie for largest probability

以上给出了回代判别出错的四个样品的编号以及它们原先属于哪一类，后来被判错的类等有关信息。

| Classification Summary for Calibration Data: WORK. IRIS | | | | | |
|--|---------|------------|-----------|--------|--------|
| Resubstitution Summary using 6 Nearest Neighbors | | | | | |
| Number of Observations and Percent Classified into Species | | | | | |
| From Species | Setosa | Versicolor | Virginica | Other | Total |
| Setosa | 50 | 0 | 0 | 0 | 50 |
| | 100.00 | 0.00 | 0.00 | 0.00 | 100.00 |
| Versicolor | 0 | 47 | 1 | 2 | 50 |
| | 0.00 | 94.00 | 2.00 | 4.00 | 100.00 |
| Virginica | 0 | 1 | 49 | 0 | 50 |
| | 0.00 | 2.00 | 98.00 | 0.00 | 100.00 |
| Total | 50 | 48 | 50 | 2 | 150 |
| | 33.33 | 32.00 | 33.33 | 1.33 | 100.00 |
| Priors | 0.33333 | 0.33333 | 0.33333 | | |
| Error Count Estimates for Species | | | | | |
| | Setosa | Versicolor | Virginica | Total | |
| Rate | 0.0000 | 0.0600 | 0.0200 | 0.0267 | |
| Priors | 0.3333 | 0.3333 | 0.3333 | | |

以上给出了全部样品回代判别的结果，错判率为 2.67%。

| Classification Results for Calibration Data: WORK. IRIS | | | | | | |
|---|--------------|-------------------------|---|--------|------------|-----------|
| Cross-validation Results using 6 Nearest Neighbors | | | | | | |
| Posterior Probability of Membership in Species | | | | | | |
| Obs | From Species | Classified into Species | | Setosa | Versicolor | Virginica |
| 84 | Versicolor | Virginica | * | 0.0000 | 0.0000 | 1.0000 |
| 134 | Virginica | Versicolor | * | 0.0000 | 0.8305 | 0.1695 |

* Misclassified observation

以上给出了用交叉验证法判错的两个样品编号及其错判情况。

| Number of Observations and Percent Classified into Species | | | | |
|--|---------|------------|-----------|--------|
| From Species | Setosa | Versicolor | Virginica | Total |
| Setosa | 50 | 0 | 0 | 50 |
| | 100.00 | 0.00 | 0.00 | 100.00 |
| Versicolor | 0 | 49 | 1 | 50 |
| | 0.00 | 98.00 | 2.00 | 100.00 |
| Virginica | 0 | 1 | 49 | 50 |
| | 0.00 | 2.00 | 98.00 | 100.00 |
| Total | 50 | 50 | 50 | 150 |
| | 33.33 | 33.33 | 33.33 | 100.00 |
| Priors | 0.33333 | 0.33333 | 0.33333 | |

| Error Count Estimates for Species | | | | |
|-----------------------------------|--------|------------|-----------|--------|
| | Setosa | Versicolor | Virginica | Total |
| Rate | 0.0000 | 0.0200 | 0.0200 | 0.0133 |
| Priors | 0.3333 | 0.3333 | 0.3333 | |

以上给出了交叉验证法判别的结果，其误判率为 1.33%。

| Obs | SepalLength | SepalWidth | PetalLength | PetalWidth | Setosa | Versicolor | Virginica |
|-----|-------------|------------|-------------|------------|---------|------------|-----------|
| 1 | 6.3 | 3.0 | 5.8 | 2.1 | 0.00000 | 0 | 0.44500 |
| 2 | 4.7 | 3.6 | 1.4 | 0.3 | 1.64276 | 0 | 0.00000 |

由此输出结果可知，新个体 A 被判为第三类 (Virginica)、新个体 B 被判为第一类 (Setosa)。注意：采用非参数判别分析，通常没有相应的判别函数式，取而代之的是贝叶斯公式。

30.4 本章小结

本章介绍了适于进行判别分析的两种数据结构之一，即原因变量为定量变量、结果变量为定性变量之情形。适于处理定量资料的判别分析方法比较多，其中又分为参数法和非参数法。

参数法通常要求定量资料服从多元正态分布，最好类与类之间方差和协方差矩阵相等。但因 SAS 软件中并没有提供检验定量资料是否服从多元正态分布的方法，故只能假定该条件满足。在多元正态分布的假定成立的前提下，若方差和协方差矩阵相等，则将各类资料合并求方差和协方差矩阵，进而计算线性判别函数的系数；否则，采用各类资料的方差和协方差矩阵计算二次型判别函数的系数。

有时，非参数判别法的效果很好，它可以通过改变其待定系数的值，来获得比较满意的判别结果。例如，可以改变 K 最近邻判别法中的 K 值或改变核密度判别法中的两个选项，即改变 R 值或/和核密度的形式 (共有正态核密度、均匀核密度等 5 种选项)。

在采用参数法或非参数法对定量资料进行判别分析之前，最好采用逐步判别分析方法进行变量筛选，以便淘汰掉无区分能力的定量变量，有利于提高判别的效果。而且，对于原因变量为定量变量的判别分析，宜先采用逐步判别分析法筛选变量，再对保留下来的定量变量采用参数法和非参数法进行判别分析，并尽可能将可变系数或选项取遍各种可能的情况，从中选择使交叉验证的误判率最低的方法。

SAS 中的典型判别分析并不太常用，因为它没有提供回代判别和交叉验证的结果，只能借用 FREQ 过程间接实现回代判别。另外，尽量不要采用多重 logistic 回归分析来间接实现判别分析，因为常得不到方程中参数的最大似然估计值，有时即便得到了参数估计值，也只能用 FREQ 过程得到回代判别结果，没有交叉验证结果，而且一般情况下误判率比较高；虽然用二值线性回归分析间接实现判别分析比较简单，但其误判率一般也比较高，且只能用 FREQ 过程间接实现回代判别。

(胡良平 刘惠刚 李子建)

第 31 章 原因变量为定性变量的判别分析

第 30 章中详细介绍了原因变量是定量变量的判别分析方法，本章将主要介绍原因变量是定性变量的判别分析方法。

31.1 实 例

31.1.1 问题与数据

【例 31-1】 某医院将 702 例阑尾炎病患者分成 3 类：慢性(π_1 ，简记为甲)、急性已穿孔(π_2 ，简记为乙)、急性未穿孔(π_3 ，简记为丙)。共考察了 7 项指标，每项指标又分若干细目。根据患者各种有关的临床症状出现频率，将资料整理成表 31-1 的形式。研究者希望根据此资料，寻找数据内部已经存在的某种规律，以便对阑尾炎的具体分类做出判断。设现有一病例各项症状及其细目如下：

上腹开始痛，恶心(+)呕吐(-)，正常便，仅右下腹部有压痛，肌紧张(-)反跳痛(+)，体温 37.3℃，白细胞 7200，这 7 种症状对应的编号依次为：3、7、9、12、15、18、20(症状编号请参阅第 31.3.2 节的表 31-2 内容)。请判断该患者患的阑尾炎最可能属于哪一种类型。

表 31-1 各类阑尾炎在各细目上出现的频率

| 症 状 | 症状的细目 | 频率(%) | | |
|------------------|--|-------|----|----|
| | | 型别：甲 | 乙 | 丙 |
| ①腹痛开始部位 X_1 ： | 右下腹 | 67 | 17 | 11 |
| | 下腹 | 2 | 4 | 5 |
| | 上腹 | 15 | 29 | 42 |
| | 脐周围 | 12 | 38 | 26 |
| | 全腹 | 5 | 11 | 15 |
| ②恶心呕吐 X_2 ： | 恶心(-)呕吐(-) | 33 | 21 | 12 |
| | 恶心(+)呕吐(-) | 53 | 39 | 28 |
| | 恶心(+)呕吐(+) | 15 | 40 | 60 |
| ③大便状况 X_3 ： | 正常便 | 86 | 74 | 53 |
| | 非正常便 | 11 | 13 | 25 |
| | 腹泻 | 3 | 13 | 22 |
| ④压痛范围 X_4 ： | 仅右下腹 | 98 | 91 | 61 |
| | 大于右下腹 | 2 | 9 | 39 |
| ⑤肌紧张和反跳痛 X_5 ： | 肌紧张(+) | 10 | 57 | 92 |
| | 肌紧张(-)反跳痛(+) | 37 | 32 | 4 |
| | 肌紧张(-)反跳痛(-) | 52 | 11 | 4 |
| ⑥体温 X_6 ： | $X_6 \leq 37^\circ\text{C}$ | 70 | 29 | 9 |
| | $37^\circ\text{C} < X_6 \leq 38^\circ\text{C}$ | 27 | 54 | 32 |
| | $X_6 > 38^\circ\text{C}$ | 3 | 17 | 59 |
| ⑦白血球含量 X_7 ： | $X_7 \leq 10000$ | 70 | 9 | 28 |
| | $10000 < X_7 \leq 15000$ | 20 | 41 | 28 |
| | $X_7 > 15000$ | 10 | 50 | 56 |

31.1.2 对数据结构的分析

表 31-1 所呈现的数据结构在很多期刊里都十分常见,但却不太符合统计表的制表要求或规范。表 31-1 实际上是对资料的一种压缩的表达形式,它最主要的特点是:在同一个表中表达了同一批受试对象多个方面的情况。本例中,受试对象为三种类型的阑尾炎患者,对每一种类型的阑尾炎患者,都分别根据腹痛开始部位、恶心呕吐、大便状况、压痛范围、肌紧张和反跳痛、体温、白血球含量 7 个方面症状中的每一种症状来进行分类,将患者分到每一种症状的不同水平(即症状细目)中去,看某一种类型的阑尾炎患者按某一种症状的不同细目分类,分到各症状细目的患者所占的频率是多少。比如,乙类阑尾炎患者按腹痛开始部位分类,腹痛开始部位为右下腹的占 17%,为下腹的占 4%,为上腹的占 29%,为脐周围的占 38%,为全腹的占 11%,其他依此类推。如此,即可读懂表 31-1 中数据的含义。

31.1.3 分析目的与统计分析方法的选择

对这种数据结构资料,若研究目的仅仅是为了筛选对结果的影响有统计学意义的变量并对结果进行预测,很可能会选择“结果变量为多值名义变量的多重 logistic 回归分析(常称为扩展的多重 logistic 回归分析)”,此时需将表中的频率转换成频数;但若研究目的是为了对新患者进行此种疾病的鉴别或分型,更常见的做法是基于最大似然法进行定性资料(指原因变量)的判别分析。

31.2 原因变量为定性变量的判别分析简介

结果变量是定性的、原因变量也是定性的,对这类资料进行判别分析时,所用的方法称为原因变量为定性变量的判别分析。适于处理这种资料的判别分析方法很少,一般只有最大似然判别法和贝叶斯公式判别法。

31.3 原因变量为定性变量的判别分析

31.3.1 SAS 程序

对于例 31-1 资料,可以运用下面的 SAS 程序获得各种症状在各细目上的得分值,其程序名为 SASTJFX31_1.SAS。有了得分值之后,再运行 SASTJFX31_2.SAS 可实现对给定的病例的类别归属进行判别。

SASTJFX31_1.SAS 的程序语句如下:

| | |
|---|--|
| <pre>DATA qualitativedata_da; infile 'd:\sastjfx\ssg101.dat'; %LET v=7; %LET m=3; ARRAY c(&v); ARRAY y(&v,&v,&v); c(1)=5; c(2)=3; c(3)=3; c(4)=2; c(5)=3; c(6)=3; c(7)=3;</pre> | <pre>FILE PRINT; DO i=1 TO &v; DO k=1 TO c(i); PUT @ 10 'x(' i ',' k ') ' @ 25 y(i,k,1) @ 35 y(i,k,2) @ 45 y(i,k,3); END; PUT;</pre> |
|---|--|

| | |
|--|--|
| <pre> DO i =1 TO &v; DO k =1 TO c(i); DO j =1 TO &m; INPUT a @@ ; b =10* (log10 (a/100) +1); y(i,k,j) =round(b,1); OUTPUT; END; END; END; ODS HTML; </pre> | <pre> END; RUN; ODS HTML CLOSE; </pre> |
|--|--|

【程序说明】 将表 31-1 中的 22 行 3 列数据输入计算机，创建一个名为 ssg101. dat 的文本格式文件，存放在 D 盘 sastjfx 文件夹内。程序中定义了两个宏变量，其中 V =7 代表症状个数，m =3 代表阑尾炎分类数，这两个数需要根据拟解决的不同问题而做相应调整。

根据具体情况，还需要调整的数据是各种症状之下的细目数，它们对应的语句是：

| |
|---|
| <pre> c(1) =5; c(2) =3; c(3) =3; c(4) =2; c(5) =3; c(6) =3; C(7) =3; </pre> |
|---|

其含义是：症状 1 有 5 个细目；症状 2 有 3 个细目；…；症状 7 有 3 个细目。

对给定病例，判断其所患的阑尾炎最可能属于哪一种类型，所需要的 SAS 程序如下，程序名为 SASTJFX31_2. SAS。

| | |
|---|---|
| <pre> data s_data; do r =1 to 22; input jia yi bing @@ ; if r =3 or r =7 or r =9 or r =12 or r =15 or r =18 or r =20 then output; end; cards; 8 2 0 -7 -4 -3 2 5 6 1 6 4 -3 0 2 5 3 1 7 6 4 2 6 8 9 9 7 0 1 4 -5 1 3 10 10 8 -7 0 6 0 8 10 6 5 -4 7 0 -4 8 5 0 4 7 5 -5 2 8 8 0 4 3 6 4 0 7 7 ; run; </pre> | <pre> ods html; proc summary data =s_data; var jia yi bing; output out =sum_value sum =sum_jia sum_yi sum_bing; run; proc print data =sum_value; var sum_jia sum_yi sum_bing; run; ods html close; </pre> |
|---|---|

【程序说明】 关键要弄清待诊断的患者的 7 种症状具体细目所对应的“编号”，根据具体情况修改程序中数据步的 if 语句即可。

31.3.2 主要分析结果及解释

各种症状在各细目上的得分值依据表 31-2 计算，得如下结果：

| Obs | sum_jia | sum_yi | sum_bing |
|-----|---------|--------|----------|
| 1 | 46 | 42 | 30 |

这是对新病例的判别结果。本例甲、乙、丙三列的合计值分别为 46、42 和 30，故判断此患者所患的阑尾炎类型最可能为甲型(即此患者患的是慢性阑尾炎)。

表 31-2 分析表 31-1 资料产生的各类阑尾炎患者在各细目上的得分

| 症状及细目的编号 | | 得 分 | | |
|----------|----|-----|----|----|
| | | 甲 | 乙 | 丙 |
| x(1,1) | 1 | 8 | 2 | 0 |
| x(1,2) | 2 | -7 | -4 | -3 |
| x(1,3) | 3 | 2 | 5 | 6 |
| x(1,4) | 4 | 1 | 6 | 4 |
| x(1,5) | 5 | -3 | 0 | 2 |
| x(2,1) | 6 | 5 | 3 | 1 |
| x(2,2) | 7 | 7 | 6 | 4 |
| x(2,3) | 8 | 2 | 6 | 8 |
| x(3,1) | 9 | 9 | 9 | 7 |
| x(3,2) | 10 | 0 | 1 | 4 |
| x(3,3) | 11 | -5 | 1 | 3 |
| x(4,1) | 12 | 10 | 10 | 8 |
| x(4,2) | 13 | -7 | 0 | 6 |
| x(5,1) | 14 | 0 | 8 | 10 |
| x(5,2) | 15 | 6 | 5 | -4 |
| x(5,3) | 16 | 7 | 0 | -4 |
| x(6,1) | 17 | 8 | 5 | 0 |
| x(6,2) | 18 | 4 | 7 | 5 |
| x(6,3) | 19 | -5 | 2 | 8 |
| x(7,1) | 20 | 8 | 0 | 4 |
| x(7,2) | 21 | 3 | 6 | 4 |
| x(7,3) | 22 | 0 | 7 | 7 |

31.4 本章小结

本章介绍了适于进行判别分析的两种数据结构之一，即原因变量为定性变量、结果变量为定性变量的情形。适于处理这种资料的判别分析方法很少，一般只有最大似然判别法和贝叶斯公式判别法。

本章以实例为牵引，给出了相应的 SAS 程序，并对输出结果进行了解释，方便读者参考修改及调用。

(胡良平 李子建 刘惠刚)